Estimating Network Features and Associated Measures of Uncertainty and Their Incorporation in Network Generation and Analysis

Dissertation Advisors: Professors Victor DeGruttola and Joseph Blitzstein Ravi Goyal

Estimating Network Features and Associated Measures of Uncertainty and Their Incorporation in Network Generation and Analysis

Abstract

The efficacy of interventions to control HIV spread depends upon many features of the communities where they are implemented, including not only prevalence, incidence, and per contact risk of transmission, but also properties of the sexual or transmission network. For this reason, HIV epidemic models have to take into account network properties including degree distribution and mixing patterns. The use of sampled data to estimate properties of a network is a common practice; however, current network generation methods do not account for the uncertainty in the estimates due to sampling.

In chapter 1, we present a framework for constructing collections of networks using sampled data collected from ego-centric surveys. The constructed networks not only target estimates for density, degree distributions and mixing frequencies, but also incorporate the uncertainty due to sampling. Our method is applied to the National Longitudinal Study of Adolescent Health and considers two sampling procedures. We demonstrate how a collection of constructed networks using the proposed methods are useful in investigating variation in unobserved network topology, and therefore also insightful for studying processes that operate on networks.

In chapter 2, we focus on the degree to which impact of concurrency on HIV incidence in a community may be overshadowed by differences in unobserved, but local, network properties. Our results demonstrate that even after controlling for cumulative ego-centric properties, i.e. degree distribution and concurrency, other network properties, which include degree mixing and clustering, can be very influential on the size of the potential epidemic.

In chapter 3, we demonstrate the need to incorporate information about degree mixing patterns in such modeling. We present a procedure to construct collections of bipartite networks, given point estimates for degree distribution, that either makes use of infor-

mation on the degree mixing matrix or assumes that no such information is available. These methods permit a demonstration of the differences between these two network collections, even when degree sequence is fixed. Methods are also developed to estimate degree mixing patterns, given a point estimate for the degree distribution.

Contents

Acknowledgments

I was fortunate to have two advisors. Victor is a constant inspiration to me. He dreams big and works hard to make his amazing visions a reality. I am honored to have had the opportunity to observe someone so passionate in creating real-world change. Joe is a mastermind of networks. He knows them inside and out and constantly directed me on the geodesic path from the question to the answer. I would also like to thank my committee member, Ted, who has an amazing ability to digest complex issues and boil them down to the comprehensible.

Thanks to Dan and Melanie, my (expanding) Cambridge family. They are always there to celebrate the holidays and life's triumphs, and to provide comfort and support through the rough patches.

Jen and Nathan, thank you for working my mind and my liver during the endless hours at Lamont and Charlie's. Jen, thank you for helping me navigate the four years of graduate school.

I am indebted to my parents who sacrificed so much for me - even too much. I am always in awe of their giving. Thanks go out to my brother, who paved a wide path for me to follow. I owe much of my early success to him. He allowed me to follow until I felt comfortable to start my own path. I will also be honored to be called 'little Raj'.

And to my wife. She has brightened every aspect of my life. During any challenge she has occasionally stood by my side, but more often has pushed me forward. Thank you for expanding my mind, filling my spirit, and opening my heart. A lot has changed since we met fourteen years ago when I wore blue mesh shorts, a white undershirt, a straw hat, flip-flops, and a scruffy beard - though I still have the beard - but your love, support and kindness has always been a constant. Thank you.

Network Construction for Ego-Centric Data

Ravi Goyal, Joseph Blitzstein, and Victor DeGruttola
Department of Biostatistics
Harvard School of Public Health

1.1 Introduction

Collecting network information from surveys is challenging, especially when the information collected is personal, like that regarding sexual behavior. As a result, many network studies collect only ego-centric data, i.e. attributes about the respondents and their contacts. Additionally, the population of interest is often too large for complete census, making it feasible only to sample small fractions of individuals; though, in many settings, researchers require an understanding of the complete network. Therefore, interest lies not only in estimating network properties, but also in constructing a robust collection of networks that have properties similar to those estimated for the unknown network under investigation. We present a novel approach to construct networks that target estimates and variances associated with the estimates for density, degree distributions and mixing patterns–properties that are estimable from ego-centric survey data.

Sampled network data allow investigators to estimate means of population characteristics, but knowledge of the complete network structure is required for investigators to simulate processes operating on networks. For example, using a complete network, an investigator can model both the diffusion of diseases or behaviors within a population and the effects of interventions to reduce the intensity of such processes. A collection of networks allows for simulations to be performed on many probable realizations of the population for which the network is partially observed, and thereby allows for characterization of the reliability of the conclusion. Even in networks for which observations are intended to be complete, collections are necessary because networks evolve over time and are often measured with errors and uncertainties. Examples of research utilizing network collections include investigation of disease control strategies for Mycoplasma pneumoniae in hospitals (Bansal et al., 2006), influenza vaccination programs within an urban population (Meyers et al., 2003), and management of tuberculosis progression within an HIV infected population (Mills et al., 2011). Collections of networks have also been used to study factors that account for differences in prevalence of sexually transmitted diseases among groups (Morris et al., 2009) and the benefit of test and treat strategies to control HIV in Sub-Saharan communities (Palombi et al., 2012).

Relationships among network properties tend to exhibit sharp threshold effects causing joint distributions to be peaked (Newman, 2010); therefore, modeling uncertainty in network properties is crucially important in network science. Erdős and Rényi (Erdős and Rényi, 1960) originally demonstrated such threshold phenomena by analytically relating the size of the largest component of an Erdős-Rényi graph to its expected mean degree. Figure 1.1, which depicts this relationship, demonstrates that failure to account properly for uncertainty can lead to misconceptions regarding processes operating on a network of interest. For the two properties they considered, this is especially a concern near the threshold where the expected node degree is one. Subsequent research has shown threshold behavior in the relationship between mean degree and categorization of networks as connected, Hamiltonian, or planar as well as size of the largest clique and the network diameter under the Erdős-Rényi random graph model (Friedgut and Kalai, 1996). Since these threshold values vary and may depend on other estimated properties, parameter uncertainty must be considered in generating networks and in using the resulting models in research.

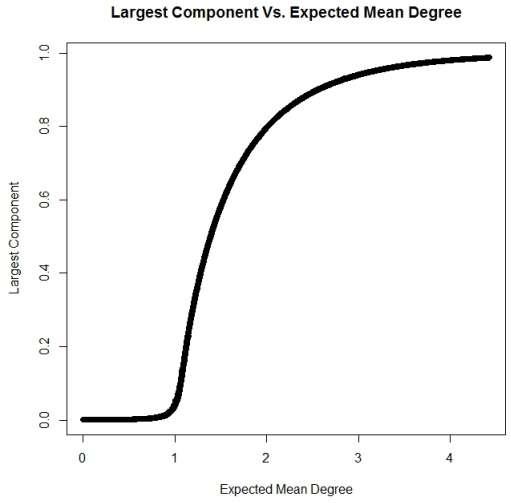

Figure 1.1: Relationship between percentage of nodes in the largest component and expected mean degree under the Erdős-Rényi random graph model (n=10,000). Note threshold at mean degree of one.

There are two additional reasons for focusing on degree distribution and mixing pat-

terns beyond the relative ease of collecting relevant data for their estimation. First, the degree distribution and mixing patterns have been shown to be sufficient to reconstruct networks for many settings (Mahadevan et al., 2006); and second, these features have great influence on processes operating on networks. The importance of mixing patterns, including age, social position, geographic location and race, has been studied in many settings, in particular sexual disease transmission (Morris et al., 2007). Degree mixing has been of particular focus in a variety of areas of research including investigation of disease transmission models (Newman, 2002), the Internet (Doyle et al., 2005; Vázquez et al., 2002), and biological interactions (Maslov and Sneppen, 2002). Newman (2002) concluded that degree assortative networks disseminate disease more easily and are more robust to removal of their highest degree nodes compared to disassortative networks.

Considerable attention has been paid to generation of networks with particular types of degree distribution (Erdős and Rényi, 1960; Barabasi and Albert, 1999; Molloy and Reed, 1995), but these efforts do not incorporate uncertainty in the degree distribution due to sampling. The Exponential Random Graph Model – a flexible approach to modeling a wide range of network properties (Frank and Strauss, 1986) – targets estimated means, but fixes the variability by maximizing the entropy. We present a novel method for network construction based on estimates of density, degree distributions and mixing patterns that incorporates uncertainty arising from sampling.

The next section provides a general description of the proposed method for constructing network collections. Section 1.4 provides additional detail for categories of network collections that are of general interest. Section 1.5 present results using data from The National Longitudinal Study of Adolescent Health (Add Health); and section 1.6 discusses our methods and suggests future research directions.

1.2 Network Collection Properties

To describe the method for constructing network collections requires defining terminology and notation. Let vector $D(g)$ denote the degree distribution of a graph g, where the i^{th} entry of $D(g)$, $D_i(g)$, is the percent of nodes with degree $i-1$. Let $d(g)$ represent the degree

sequence of network g, where the i^{th} entry, $d_i(g)$, is the degree of node i. Let $MM(g)$ be a matrix representing the mixing pattern of graph g. The entry $MM_{k,l}(g)$ is the percentage of edges from a node with covariate pattern k to a node with covariate pattern l. Let vector $m(g)$ represent the covariate pattern for each node in g, where $m_i(g)$ is the covariate pattern for node i. Finally, let $M(g)$ be the vector of percentages for the different covariate patterns. We use the notation $DMM(g)$ to represent degree mixing matrices, where entry $DMM_{i,j}(g)$ is the percentage of edges from a node of degree i to a node of degree j.

The network collection is a subset of networks from the space, $\mathscr{G}$, of graphs with n nodes. To construct such a collection, we begin by partitioning $\mathscr{G}$ into congruence classes, such that each network in the congruence class has the same values for network properties of interest. Let C_g represent the congruence class containing network g. For example if interest lies in degree distribution and degree mixing patterns, networks g and h will reside in the same congruence class if and only if $D(g) = D(h)$ and $DMM(g) = DMM(h)$. The congruence classes represent the finest partition of the space $\mathscr{G}$ that is based both on estimable quantities from observed data and of scientific interest. In this paper, the congruence classes will be constructed by partitioning $\mathscr{G}$ by density, degree distribution and mixing patterns. To construct a collection, each network in $\mathscr{G}$ will be assigned a probability, $P_{\mathscr{G}}(g)$, of being selected into the collection. $P_{\mathscr{G}}(g)$ is based solely on the congruence class of the network g.

By partitioning $\mathscr{G}$ into congruence classes and defining a probability mass function for the probability of sampling a network from a congruence class we can control the probability of sampling a network with particular values for network properties. As some congruence classes have vastly more networks than do others, this approach guards against over or under representing networks with particular properties due to the size of the congruence class, thus ensuring that the collection of networks is consistent with the collected data. Defining the probability of sampling networks from a congruence class allows for construction of networks that reflect both the estimated mean and uncertainty associated with the estimate– the only information available to the investigator– without requiring consideration of the complex topology of the underlying space of graphs of size n.

A Markov chain Monte Carlo (MCMC) procedure is the basis for generating a collection of

networks, $\{g_1, \cdots, g_t\}$ that satisfy the probability distribution assigned to the congruence classes. Ideally, to construct our collection, $\{g_1, \cdots, g_t\}$, we would sample, with replacement, t congruence classes $\{C_1, \cdots, C_t\}$, based on the probability distribution on the classes. For each congruence class, C_i where $i \in \{1, \cdots, t\}$, we would draw a network, g_i, such that $g_i \in C_i$. Since this procedure presents computational difficulties, we implement a Markov chain using the Metropolis-Hastings algorithm to generate the networks. For a review on MCMC methods see Robert and Casella (2004). In order to implement the Metropolis-Hastings algorithm, four aspects have to be specified: target function, proposal function, acceptance probability, and initial starting element. Many authors have described construction of an initial starting element (Blitzstein and Diaconis, 2010), so we discuss only the first three aspects below.

1.2.1 Target Function

The target function is the desired stationary distribution for the Markov chain. In our setting, the network g has a probability mass equal to the probability of the congruence class C_g divided by the number of networks in C_g, $|C_g|$, thereby ensuring that each network in C_g has the exact same probability:

$$P_{\mathscr{G}}(g) \propto \left(\frac{1}{|C_g|}\right) * P_C(C_g). \tag{1.1}$$

Due to the constraints imposed on particular network features not all values of particular network features correspond to valid networks. For example, no network can have odd values for $\sum_i D_i * i * n$, which represents two times of the number of edges in the graph. Section 4 outlines criteria to ensure that a congruence class contains at least one valid network.

1.2.2 Proposal Function

The algorithm generates the next network, g_{t+1}, in the chain by nominating a proposal network, gp_{t+1}, based only on the previous network g_t. A common method to generate a proposal network is by toggling the existence of an edge. Edge toggling requires selecting two nodes at random and either removing the edge if one exists or adding one if it does

not. The algorithm produces an irreducible Markov chain among all graphs with a fixed size; the chain also has equal forward and backward probabilities.

1.2.3 Acceptance Probability

Once a proposal network, gp_{t+1}, is generated, the Metropolis-Hastings algorithm will either accept, $g_{t+1} = gp_{t+1}$, or reject, $g_{t+1} = g_t$, the proposal. The Metropolis-Hastings acceptance probability is the following:

$$P(\text{Accept } gp_{t+1}|g_t) = \frac{P_{\mathscr{G}}(gp_{t+1})}{P_{\mathscr{G}}(g_t)} = \frac{|C_{g_t}|}{|C_{gp_{t+1}}|} * \frac{P_C(C_{gp_{t+1}})}{P_C(C_{g_t})} \tag{1.2}$$

Let $t(g, C_h)$ equal the number of elements in C_h that differ from g through toggling one edge. Let $T(C_g, C_h) = \sum_{g \in C_g} t(g, C_h)$ represent the total number of possible edge toggles for graphs in C_g to graphs in C_h. Due to symmetry induced by edge toggling $T(C_g, C_h) = T(C_h, C_g)$. Thus,

$$P(\text{Accept } gp_{t+1}|g_t) = \frac{\left(\frac{T(C_{g_t}, C_{gp_{t+1}})}{|C_{gp_{t+1}}|}\right)}{\left(\frac{T(C_{gp_{t+1}}, C_{g_t})}{|C_{g_t}|}\right)} * \frac{P_C(C_{gp_{t+1}})}{P_C(C_{g_t})} \tag{1.3}$$

By defining $f(C_g, C_h)$ as the average number of elements in C_h that are valid proposals from an element $g \in C_g$, we get the following:

$$P(\text{Accept } gp_{t+1}|g_t) = \frac{f(C_{gp_{t+1}}, C_{g_t})}{f(C_{g_t}, C_{gp_{t+1}})} * \frac{P_C(C_{gp_{t+1}})}{P_C(C_{g_t})} \tag{1.4}$$

1.3 Network Collection Construction

This section describes several common scenarios that demonstrate the capabilities of our framework to handle sampled data. Once the network space and network proposal method have been selected, only the functions $P_C(C_g)$ and $f(C_g, C_h)$ need to be specified. In all scenarios, the network space consists of all networks with a fixed number of nodes, and edge toggling is used to propose networks. As the probability mass function, P_C, on congruence classes is set by the investigator, in this section we only derive $f(C_g, C_h)$. The next section provides examples of various probability mass functions associated with

different sampling strategies. Denote $g, h \in \mathscr{G}$ as the current and proposal network, respectively. Let C_g and C_h denote the congruence classes for g and h. Let the edge, (i, j), between node i and node j be the connection that is toggled to move from g to h and back. Without loss of generality, let $(i, j) \in h$ but $(i, j) \notin g$.

1.3.1 Topological Features

Topological features of a social network provide valuable insight into how processes operate within a community. In the following section we will discuss constructing networks based on density, degree distribution and degree assortativity. Though density is an influential network property (Bollobás, 2001) it is usually collected with other network features, e.g. degree distribution, but it provides a useful example to illustrate the mechanics of our method.

Density

For density, a congruence class is set of networks with the same number of edges, since all graphs in $\mathscr{G}$ have the same number of nodes. Let $|E_g|$ denote the number of edges in graph g. Networks g_1 and g_2 are in the same congruence class if and only if $|E_{g_1}| = |E_{g_2}|$. Since $(i, j) \in h$ but $(i, j) \notin g$, $|E_h| = |E_g| + 1$. To calculate $f(C_h, C_g)$ we need to know the average number of elements in C_g that are valid proposals from any element $h \in C_h$. Since removing any edge in h will produce a graph in C_g there are exactly $|E_h|$ valid proposals in C_g from graph h, and this is true regardless of the choice of $h \in C_h$. Thus,

$$f(C_h, C_g) = |E_h| \tag{1.5}$$

To calculate $f(C_g, C_h)$, we need to know the average number of elements in C_h that are valid proposals from any element $g \in C_g$. Adding any edge in g, which does not exist, will produce a graph in C_h, hence there are exactly $\binom{n}{2} - |E_g|$ valid proposals in C_h from graph g. Again, is it true for any $g \in C_g$. Thus,

$$f(C_h, C_g) = \binom{n}{2} - |E_g| \tag{1.6}$$

The investigator can stipulate the percentage of networks in the collection with a particular

number of edges by specifying the values of $P(C_g)$. One specification is to generate a network collection following Erdős-Rényi random graph model with parameters (n, p), which can be done by setting $P_C(C_g) = p^{|E_g|} * (1-p)^{\binom{n}{2}-|E_g|} * \binom{.5*n*(n-1)}{|E_g|}$. Another specification, which is the main propose of the paper, is to define $P(C_g)$ in such a way to generate networks based on the uncertainty due to sampling.

Degree Distribution

For degree distribution, congruence classes are sets of networks with identical numbers of nodes and degree distribution. Thus, networks g_1 and g_2 are in the same congruence class if and only if $D_k(g_1) = D_k(g_2)\ \forall\ k$. As g only differs from h through a toggling of the edge (i, j) , $D_k(g) = D_k(h)$ for all k except possibly $k = d_i(g), d_j(g), d_i(h)$ and $d_j(h)$. Since the only difference between the graph g and h is edge $(i, j) \in h$ but $(i, j) \notin g$, $d_i(h) = d_i(g) + 1$ and $d_j(h) = d_j(g) + 1$. The expressions relating $D(g)$ and $D(h)$ are given below for those entries that may differ.

$$D_k(h) = \begin{cases} D_{d_i(g)}(g) - (1 + I\{d_i(g) = d_j(g)\} - I\{d_i(g) = d_j(g) + 1\})/n \text{ if } k = d_i(g) \\ D_{d_j(g)}(g) - (1 + I\{d_i(g) = d_j(g)\} - I\{d_j(g) = d_i(g) + 1\})/n \text{ if } k = d_j(g) \\ D_{d_i(g)}(g) + (1 + I\{d_i(g) = d_j(g)\} - I\{d_i(g) = d_j(g) - 1\})/n \text{ if } k = d_i(g) + 1 \\ D_{d_j(g)}(g) + (1 + I\{d_i(g) = d_j(g)\} - I\{d_j(g) = d_i(g) - 1\})/n \text{ if } k = d_j(g) + 1 \end{cases} \tag{1.7}$$

The number of edge toggles from a graph $h \in C_h$ to any graph in C_g is equal to the percentage of edges in h that have endpoint degrees of $d_i(h)$ and $d_j(h)$, $DMM_{d_i(h),d_j(h)}$, multiplied by the number of edges in h, $|E_h|$. Thus, $f(C_h, C_g)$ is equal to the average of $DMM_{d_i(h),d_j(h)} * |E_h|$ over all graphs $h \in C_h$. Let $E(DMM|C_h)$ denote the expected degree mixing matrix over graph that are in C_h. Since $h' \in C_h$ if and only if $D(h') = D(h)$, $E(DMM|C_h) = E(DMM|D(h))$. Thus,

$$f(C_h, C_g) = E(DMM_{d_i(h),d_j(h)}|D(h)) * |E_h| \tag{1.8}$$

Following arguments from Newman (2002), based on the probability that a node's neighbor will have degree k is proportional to $k * D_k$ and not D_k,

$$E(DMM_{x,y}|D) \approx \frac{D_x * x * D_y * y}{.5 * (\sum_k D_k * k)^2} * \left(\frac{1}{2}\right)^{I\{x=y\}} \tag{1.9}$$

The number of edge toggles from a graph $g \in C_g$ to any graph in C_h is equal to the number of possible non-loop edges with endpoint degrees $d_i(g)$ and $d_j(g)$ minus the number of edges that will generate a multi-edge. The expected number of edge toggles that generate a multi-edge is $E(DMM_{d_i(g),d_j(g)}|D(g)) * |E_g|$, denote this value as α_1.

$$f(C_h, C_g) = \begin{cases} n^2 * D_{d_i(g)}(g) * D_{d_j(g)}(g) - \alpha_1 \text{ if } d_i(g) \neq d_j(g) \\ \binom{n * D_{d_i(g)}(g)}{2} - \alpha_1 \text{ else .} \end{cases} \tag{1.10}$$

To decrease convergence time in the MCMC procedure, one can use the algorithm described in Blitzstein and Diaconis (2010) to initialize the starting network with the estimated mean degree distribution. An algorithm to validate that a degree sequence has a realization can be found in Blitzstein and Diaconis (2010) which is based on results from Hakimi (1962) and Havel (1955).

Degree Mixing and Degree Distribution

We consider a partition of $\mathscr{G}$ such that networks g_1 and g_2 are in the same congruence class if and only if $D_x(g_1) = D_x(g_2)\ \forall\ x$ and $DMM_{x,y}(g_1) = DMM_{x,y}(g_2)\ \forall\ x, y$. An identical partition is defined when networks g_1 and g_2 are in the same congruence class if and only if $DMM(g_1) * |E_{g_1}| = DMM(g_2) * |E_{g_2}|$. Thus, the probability mass function can be defined using the degree mixing matrix and number of edges. Similar to degree distribution, the number of edge toggles from a graph $h \in C_h$ to any graph in C_g is equal to $f(C_h, C_g) = DMM_{d_i(h),d_j(h)} * |E_h|$, though, in this setting, $E(DMM|C_h) = DMM(h)$ because all graphs in C_h have the same DMM. Similar logic holds for $f(C_g, C_h)$, thus by substituting the true degree mixing matrix for the expected degree mixing matrix in equations (8) and (10), we get the following expressions for $f(C_g, C_h)$ and $f(C_h, C_g)$.

$$f(C_h, C_g) = DMM_{d_i(h),d_j(h)}(h) * |E_h| \tag{1.11}$$

$$f(C_g, C_h) = \begin{cases} n^2 * D_{d_i(g)}(g) * D_{d_j(g)}(g) - \alpha_2 \text{ if } d_i(g) \neq d_j(g) \\ \binom{n * D_{d_i(g)}(g)}{2} - \alpha_2 \text{ else} \end{cases} \tag{1.12}$$

where $\alpha_2 = DMM_{d_i(g),d_i(g)}(g) * |E_g|$.

As with the degree distribution, not all degree mixing matrices have a valid realization. Appendix A provides a method to characterize valid degree mixing matrices; an alternative proof of the validity of this characterization is given by Amanatidis, Green and Mihail (2008). Using the construction procedure in Appendix A to set the initial network with the estimated degree distribution and degree mixing will tend to decrease time to convergence in the MCMC procedure.

1.3.2 Nodal Covariates

The methods developed for topological network features can be extended to include mixing patterns based on nodal covariates. Let p be the number of distinct nodal covariate patterns of interest in the population. The covariate patterns can represent single or multiple nodal characteristics. We describe a common scenario in which we observe not only mixing patterns between covariate patterns but also the degree distributions, $\{D^1, \cdots, D^p\}$, for each covariate pattern. The following approach can be simplified for settings wherein individual covariate pattern degree distributions are not observed. In order to incorporate covariate information, knowledge of the percentage of individuals with covariate pattern k, M_k, is required for each k.

Nodal Covariate Mixing and Degree Distribution

For nodal covariate mixing and degree distribution, the congruence classes contain networks with identical numbers of nodes, degree distributions, and nodal covariate mixing matrices. Thus, networks g_1 and g_2 are in the same congruence class if and only if $D_x^k(g_1) = D_x^k(g_2) \forall x, k$ and $MM_{k,l}(g_1) = MM_{k,l}(g_2) \forall k, l$.

For each covariate degree distribution, one entry in the mixing matrix is fixed. Therefore, given degree distribution estimates for each of the covariate patterns, the probability mass function can only be specified for the degree distributions and the entries above the

diagonal in the mixing matrix. As above, expected degree mixing matrices, $E(DMM^{k,l})$, are constructed for each entry in the upper triangle of the covariate mixing matrix, thus $k \neq l$. The matrix entry $DMM^{k,l}_{x,y}(g)$ represents the percentage of edges where one endpoint node has covariate pattern k and degree x, while the other endpoint node has covariate pattern l and degree y. Using the setup from the previous section, we let the edge set of g and h be identical except that $(i, j) \notin g$ and $(i, j) \in h$. Regarding covariate information, let nodes i and j have covariate patterns m_i and m_j, respectively. The number of edge toggles from a graph $h \in C_h$ to any graph in C_g is equal to the number of edges in h where one endpoint has degree $d_i(h)$ and type m_i and the other endpoint has degree $d_j(h)$ and type m_j. The proportion of edges where both endpoints are specified as type m_i and type m_j compared to edges where one endpoint is required to be of type m_i is $\frac{MM_{i,j}}{MM_{i,i}+\sum_z MM_{i,z}}$. Using similar arguments as above we can calculate the expected degree mixing matrix where only edges between types m_i and m_j are considered.

$$E(DMM^{k,l}_{x,y}|D^{k,l}, D^{l,k}) \approx \frac{D^{k,l}_x * x * D^{l,k}_y * y}{(\sum_z D^{k,l}_z * z)^2} \tag{1.13}$$

where,

$$D^{k,l} = M_k * D^k * \frac{MM_{k,l}}{MM_{k,k} + \sum_z MM_{k,z}} \tag{1.14}$$

Thus,

$$f(C_h, C_g) = E(DMM^{m_i,m_j}_{d_i(h),d_j(h)}(h)|D^{m_i,m_j}(h), D^{m_j,m_i}(h)) * |E^{k,l}_h| \tag{1.15}$$

and,

$$f(C_g, C_h) = n^2 * M_{m_i} * D_{d_i(g)}(g) * M_{m_j} * D_{d_j(g)}(g) - \alpha_3 \tag{1.16}$$

where $|E^{k,l}_h| = \sum_z D^{k,l}_z * z$, the expected number of edges between the nodes types, and $\alpha_3 = E(DMM^{m_i,m_j}_{d_i(g),d_j(g)}(g)|D^{m_i,m_j}(g), D^{m_j,m_i}(g)) * |E^{k,l}_g|$.

Nodal Covariate Mixing, Degree Mixing, and Degree Distribution

In a similar fashion as above the proposed method can be extended to include degree mixing. Once again, we substitute the true degree mixing matrices for the expected degree mixing matrices.

1.4 Results

1.4.1 Add Health Data

The National Longitudinal Study of Adolescent Health (Add Health) is a longitudinal study of a nationally representative sample of adolescents in grades 7-12 in the United States during the 1994-95 school year (Harris and Udry, 2012).

The data arose from a two-stage cluster design: the first stage developed a stratified, random sample of all high schools in the United States, and the second sampled a set of students from each school. Only the publicly available data on approximately 6000 students in 132 school were used for all analysis. The data were developed from a questionnaire in which students were asked to name at most five male and five female friends. For each school $s \in \{1, \cdots, 132\}$, we estimated the overall degree distribution, $D(s)$, the degree distribution for each gender, $D^{male}(s)$ and $D^{female}(s)$, and the percent of mixing between genders, $MM(s)$.

The ADD health data permit investigation of the proposed methods in two important settings. In the first, researchers model the network for a specific community using data collected on similar communities. In this setting, estimates of uncertainty in network properties of the community under investigation are based on the variation in network properties across the subset of similar communities. In the second, data are collected on a sample from the population of interest.

In order to better illustrate the importance of accommodating uncertainty, we compare our method to two currently available approaches. The first is based on the Exponential Random Graph Model which targets estimate means, but fixes the variability by maximizing the entropy. The second fixes the network properties at the mean estimates and does not allow for variability.

1.4.2 Multiple Network Summary

In this section we generate a 'typical' student friendship network based only on the degree distribution for all 132 schools; and in the next, we consider an additional network feature, mixing pattern. We start with degree distribution alone as it is often the most important network property influencing diffusion on networks. The estimated degree distributions for the 132 schools are taken as the true degree distributions for clarity, though an investigator can incorporate the additional uncertainty arising from the estimated degree distributions.[1] Since the degree distributions from the 132 schools vary, there is no definitive degree distribution of a 'typical' school. Some degree distributions are more plausible than others, for example a school with no friendships, $D = (1, 0, 0, \cdots, 0)$, would be highly unusual. Therefore, we want to assign probabilities to possible degree distributions based on the observed 132 schools.

A Dirichlet distribution, $Dir(\alpha)$, is used to model the degree distribution, D, of a 'typical' student friendship network, since the sum of the degree distribution totals one. Table 1.1 summarizes the mean percentages for degree 0 to degree 10 using the maximum likelihood estimates (MLEs) for α.

Table 1.1: **Distribution of Number of Friendships**

Degree	Percentage
0	1.83
1	9.31
2	9.66
3	11.00
4	12.72
5	15.09
6	8.83
7	10.85
8	10.82
9	6.08
10	3.80

[1] In the publicly available dataset there are missing data, resulting in some degree distributions being quite sparse. Also not all friendships reported by the student were contained in the same school.

Using the Dirichlet distribution as the probability mass function on the congruence classes, our target function is

$$P_{\mathscr{G}}(g) \propto \left(\frac{1}{|C_g|}\right) * P_C(D = D(g)). \tag{1.17}$$

where $D \sim Dir(\alpha)$. The congruence classes are defined solely by degree distribution, two networks g_1 and g_2 are in the same congruence class if and only if $D(g_1) = D(g_2)$. We constructed 1,000,000 networks of size 100, in which the first 30,000 were removed for burn-in. For each degree $i \in \{0, \cdots, 10\}$, figure 1.2 depicts the marginal probability function, $P(D_i = D_i(g))$, that a constructed network, g, will have D_i percentage of nodes with degree i. In each of the plots, the black line is the targeted marginal density of the Dirichlet distribution. The red line is the marginal density from the algorithm presented in this paper. In each individual degree plot in figure 1.2, our method closely resembles the target distribution, thus models the estimated mean and specified variability. The blue and green lines are output from the exponential random graph model (Handcock et al., 2012) and fixed degree sequence where both use the mean estimates from table 1.1.

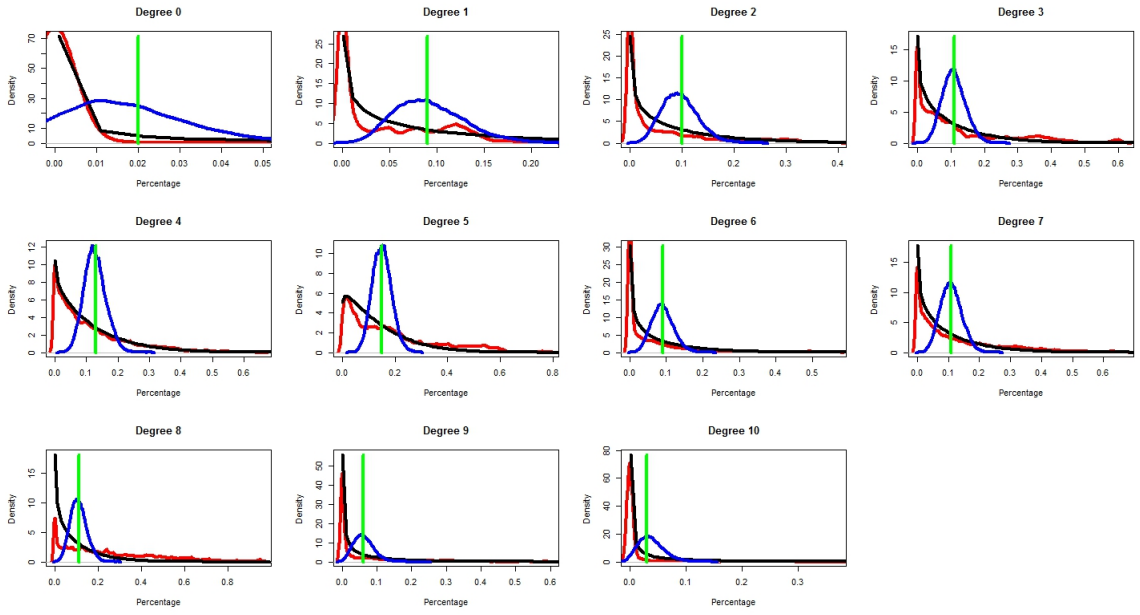

Figure 1.2: Black, red, blue and green lines represent target distribution and output from our method, ERGMs, and zero variability

To demonstrate the impact of constructing networks based on these three methods, we calculate global network properties using the igraph library (Csardi and Nepusz, 2006) in R (R Development Core Team, 2011). The first four properties represents summary measures for node centrality. The betweenness of a node is defined by the number of geodesics (shortest paths) going through the node. The closeness of a node is the inverse average distance to all other nodes in the graph. The median and max betweenness (closeness) are defined over the values of betweenness (closeness) calculated for each node. Another measure is diameter, which is defined as the length of the longest geodesic. The remaining three measures are the mean, median and max component size.

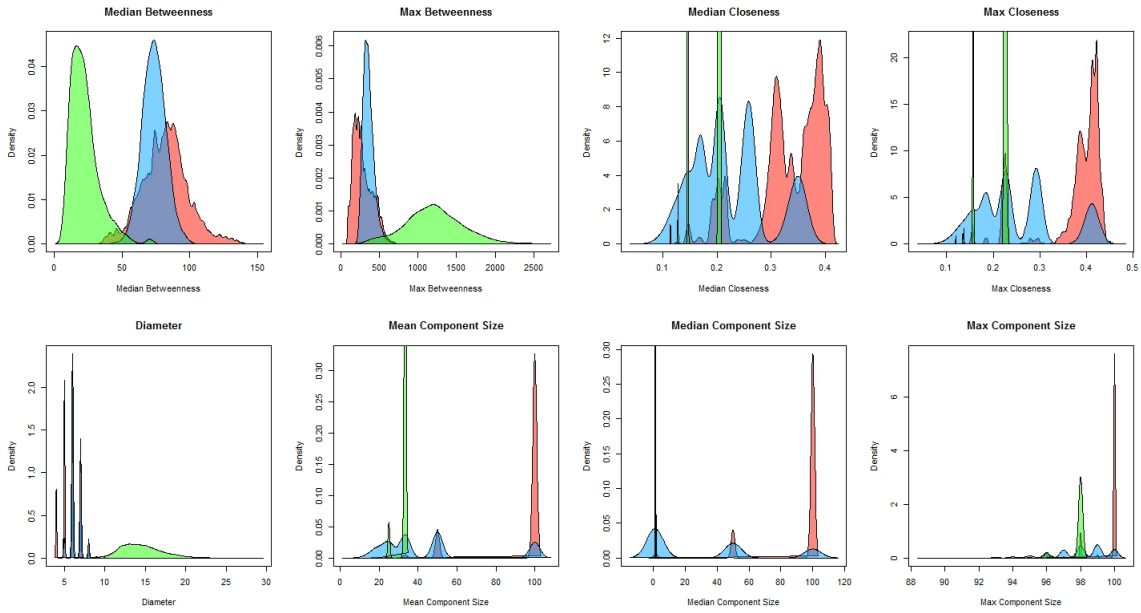

Figure 1.3: Red, blue, and green densities represent global properties from our algorithm, ERGMs, and fixed degree respectively

Figure 1.3 shows that the distributions of many global properties differ considerably among the three approaches, highlighting the importance of accurate quantification of uncertainty of estimated network properties. Different specifications of the uncertainty in microscopic node behavior can lead to dramatic differences in the macroscopic network structure. Thus, care should be taken in specifying the method to generate the network collection to ensure that the reliability of the conclusion is depicted accurately.

1.4.3 Single Sampled Network

Some common ego-centric network study designs collect information that permits estimation of both degree distribution and non-degree mixing patterns. Collecting information on degree mixing patterns can be challenging, as it typically requires a form of link tracing. The example below represents a scenario in which interest lies in understanding the population from which the sample was drawn. We investigate one school from the ADD Health survey in which 32 males and 47 females provided information on their number of friendships (degree), genders, and the genders of their friends. Information of gender mixing was provided on 418 ties. Tables 1.2 and 1.3 list mean percentages for both male and female degree distributions and gender mixing.

Table 1.2: **Degree Distribution by Gender**

Degree	Female	Male
0	0.00	0.00
1	0.04	0.12
2	0.09	0.16
3	0.15	0.16
4	0.11	0.03
5	0.06	0.09
6	0.15	0.06
7	0.15	0.16
8	0.13	0.06
9	0.09	0.12
10	0.04	0.03

Table 1.3: **Percentage of Mixing Between Genders**

	Male	Female
Male	.239	.400
Female	.400	.361

It is reasonable to consider that the number of friendships for each student are drawn from a multinomial distribution, thus we can estimate the covariance matrices using a multivariate normal approximation. The (D_i^k, D_j^k) entry in the covariance matrix for each

degree distribution, $k \in \{male, female\}$, is given by $\frac{D_i^k * (1 - D_i^k)}{X^k}$ if $i = j$ and $\frac{-D_i^k * D_j^k}{X^k}$ otherwise where $X^{male} = 32$ and $X^{female} = 47$. The male-female gender mixing distribution is assumed to approximate a normal distribution, $N(\hat{\mu} = .400, \hat{\sigma}^2 = \frac{(.400) * (1 - .400)}{418})$. The other gender mixing categories, male-male and female-female, are fixed once the degree distributions and male-female mixing are specified. We construct 1,000,000 networks (30,0000 removed for burn-in) of 100 individuals where the number of individuals that are male (41%) or female (59%) are proportion to those in the sample. Figures 1.4- 1.6 show results for both degree distributions and mixing pattern. Again, the black line is the target density, while the red and blue lines represent output from our method and the Exponential Random Graph Model, respectively. In this example, ERGM aligns close to the target density, though there does exist divergence, particularly for the gender mixing distribution. As seen in figure 1.7, our method and ERGMs produce networks with similar global properties.

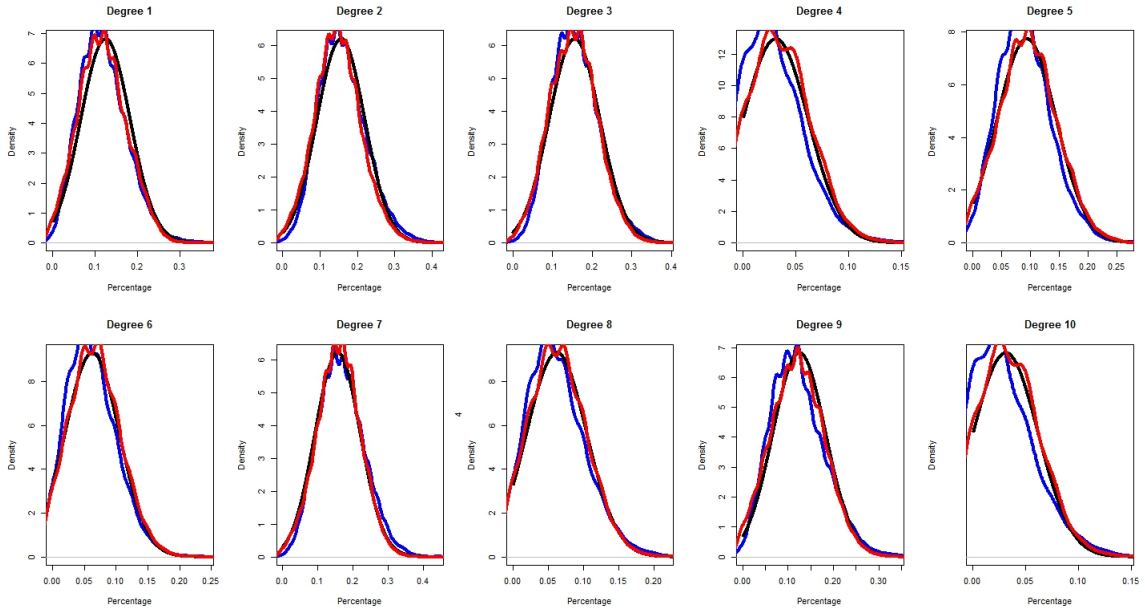

Figure 1.4: Black, red and blue lines represent target distribution and output from our method and ERGMs.

Table 1.4 gives 95% confidence intervals for both the targeted and the constructed degree and mixing distributions for our method.

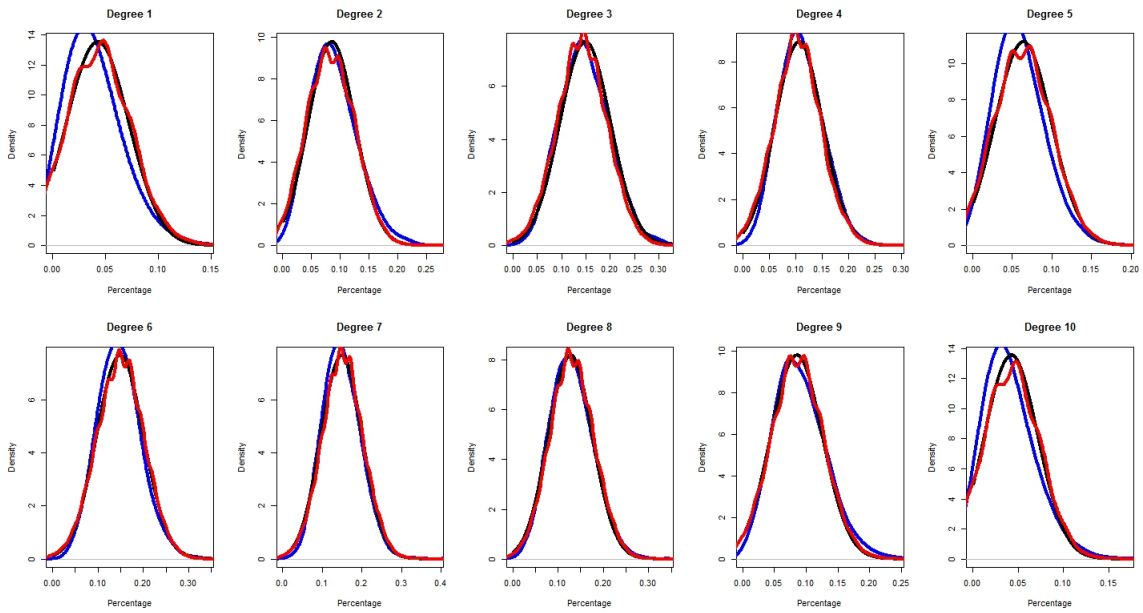

Figure 1.5: Black, red and blue lines represent the target distribution results from the proposed method and from ERGMs, respectively.

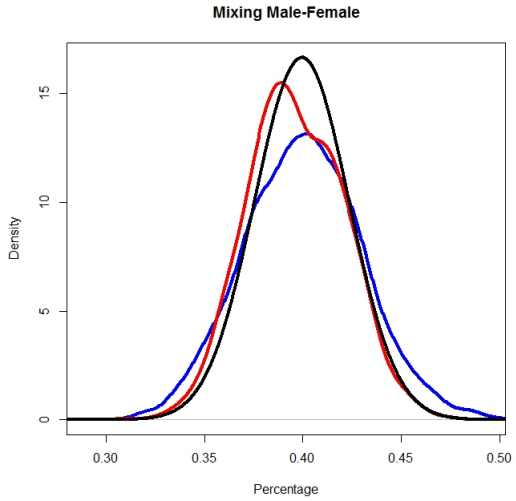

Figure 1.6: Black, red and blue lines represent the target distribution results from the proposed method and from ERGMs, respectively.

Though the method described in this paper matches the targeted distribution closely, care still needs to be used when interpreting the collection of networks that are constructed using this method. The probability function applies to the congruence classes, and not all

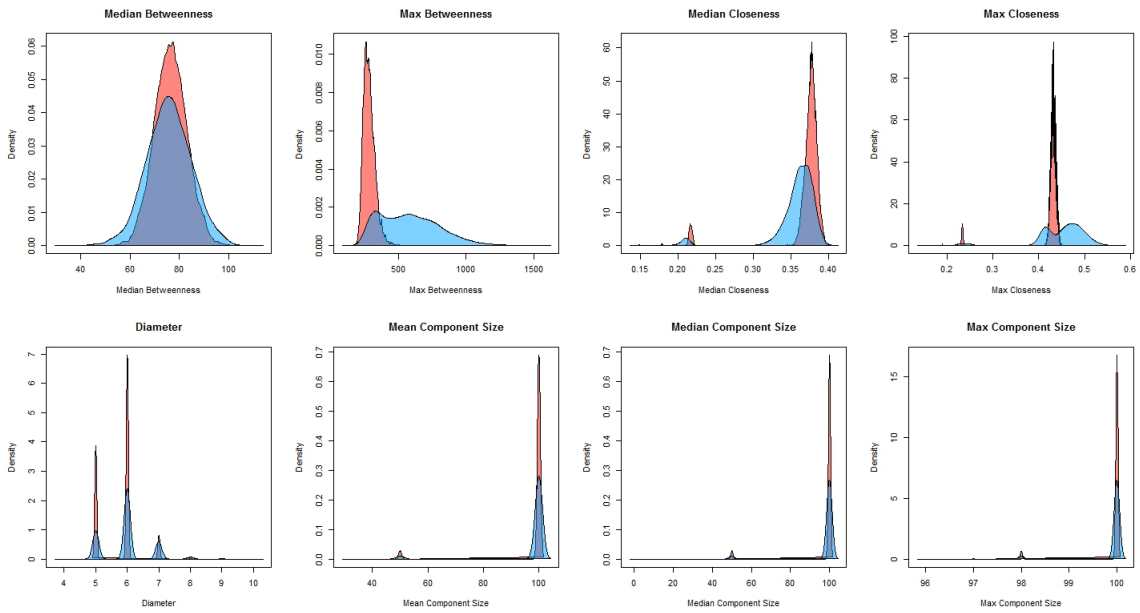

Figure 1.7: Red and blue densities represent global properties from our algorithm and ERGMs, respectively.

degree distributions and mixing matrices have valid networks associated with them.

1.5 Discussion

This paper presents novel methods to incorporate uncertainty due to sampling in the construction of networks. The network properties of density, degree distribution and mixing patterns were considered in illustration of the approach. Degree distribution and mixing patterns have been shown to have great influence on processes operating on diverse areas such as Internet connectivity, biological interactions and sexual disease transmission.

The proposed methods allowed construction of collections of networks that take into account uncertainty that arises from sampling, but the methods can also be used to incorporate uncertainty that results from reporting errors, which may appear even in fairly simple sampling designs. Hence, these methods are well suited to model propagation of infectious diseases, especially sexually transmitted infections, on networks.

The ability to accommodate uncertainty in estimated network properties in the construction of collections of networks allows investigators to assess the level of precision in

Table 1.4: **Comparison Between Target and Generated Networks**

		Constructed	Target
Category	Parameter	95% CI	95% CI
Male Degree Distribution	D_0^{Male}	(0-0)	(0-0)
	D_1^{Male}	(2.4-21.9)	(2.8-22.1)
	D_2^{Male}	(4.8-24.3)	(5.0-26.1)
	D_3^{Male}	(4.8-24.3)	(5.0-26.1)
	D_4^{Male}	(0.0-7.3)	(-2.0-8.1)
	D_5^{Male}	(2.4-17.0)	(0.8-17.8)
	D_6^{Male}	(0.0-12.1)	(-0.8-13.2)
	D_7^{Male}	(4.8-26.8)	(5.0-26.1)
	D_8^{Male}	(0.0-12.1)	(-0.8-13.2)
	D_9^{Male}	(2.4-21.9)	(2.8-22.1)
	D_{10}^{Male}	(0.0-7.3)	(-2.0-8.1)
Female Degree Distribution	D_0^{Female}	(0.0-0.0)	(0.0-0.0)
	D_1^{Female}	(0.0-9.7)	(-0.6-9.0)
	D_2^{Female}	(2.4-14.6)	(1.8-15.2)
	D_3^{Female}	(4.8-21.9)	(6.3-23.4)
	D_4^{Female}	(2.4-17.0)	(3.2-18.0)
	D_5^{Female}	(0.0-12.1)	(0.5-12.2)
	D_6^{Female}	(7.3-24.3)	(6.3-23.4)
	D_7^{Female}	(7.3-24.3)	(6.3-23.4)
	D_8^{Female}	(4.8-21.9)	(4.7-20.7)
	D_9^{Female}	(2.4-14.6)	(1.8-15.2)
	D_{10}^{Female}	(0.0-9.7)	(-0.6-9.0)
Mixing	$MM_{Male,Female}$	(35.7-43.6)	(36-43.8)

these estimates that is needed for reliable evaluation of the relative merits of different policy options. Hence, these approaches can be useful in designing randomized trials of community-level control strategies in settings where infections or behaviors diffuse over social or sexual networks.

Further research is needed to further expand this framework to include additional network properties, in particular clustering. This framework can also be useful in constructing dynamic networks that allow variation over time.

Sampling Dynamic Networks to Understand Impact of Concurrency

Ravi Goyal, Joseph Blitzstein, and Victor DeGruttola

Department of Biostatistics

Harvard School of Public Health

2.1 Introduction

The presence of concurrent partnerships has been shown to increase the spread of HIV within communities in research based on theoretical models (Morris et al., 2009; Morris and Kretzschmar, 1997; Kretzschmar and Morris, 1996; Jeffrey W. Eaton and Garnett, 2011). Interest in concurrency has led to efforts to establish the best way to measure concurrency and to assess accuracy of estimates based on these measures (Lurie and Rosenthal, 2010; Morris, 2010). Nonetheless, theoretical claims have not yet been supported by empirical evidence (Tanser et al., 2011; Sawers and Stillwaggon, 2010). Proposed reasons for this discrepancy include the complexity of the challenges that arise in estimating the contribution of concurrent relationships to the risk of HIV infection among susceptible members of the community. Other explanations include differences in sexual practices between partnerships occurring simultaneously compared to sequentially, e.g. coital dilution (Sawers et al., 2011). In this paper we focus on a different potential explanation for the discrepancy, i.e. the degree to which impact of concurrency on HIV incidence in a community may be overshadowed by differences in other unobserved, but local, network properties. This paper provides insight into the assumptions about the uniformity of typically unobserved sexual network properties across communities with possibly different population compositions, geography, transportation, and cultural norms that are required (at least implicitly) to estimate the impact of concurrency on HIV incidence from empirical data. Since there is evidence that communities differ considerably in values for such observed network properties as distribution of number of partners, concurrency, and mixing patterns (Morris et al., 2007), it is plausible that communities may differ in other network properties which are not observed. In fact, there already exists indirect evidence that unobserved properties may differ between communities, for example observed levels in the practice of polygyny vary across regions (Reniers and Watkins, 2010), which can influence cumulative measurements of degree mixing - a property not typically estimable from survey data. In this paper we demonstrate that slight differences in partner selection behavior, which affect unobserved properties, offer a possible explanation for the failure to observe the effect of concurrency on actual estimates of HIV incidence in a community.

We also discuss how unobserved network measures provide insight into how certain types of concurrency, in particular polygynous relationships, can appear to be protective at the community level, also referred to as "benign concurrency" (Reniers and Watkins, 2010), but detrimental at an individual level (Kretzschmar et al., 2010).

In order to investigate the plausibility that unobserved network properties may be highly influential, we need to separate the effect of observed network properties, degree distribution and concurrency, from that of properties that are typically unobserved. This is made challenging by the fact that in many situations, changing one network property modifies others, rendering the incremental impact of any one network property difficult to ascertain. We develop a method to randomly sample dynamic networks uniformly, given both a cumulative contact network and observed measurements for cumulative concurrency. By applying the method to a set of cumulative contact networks–all with a prescribed degree distribution but variable values for an unobserved network property–we can evaluate the incremental impact of the property, beyond the effect of degree distribution and concurrency, on measures for HIV incidence. Using the same approach we are able to estimate the outcome for HIV incidence for communities with only sequential monogamous relationships by setting the metrics that characterize concurrency to zero. For both scenarios, one with concurrency and the other without, the communities have identical cumulative contact networks; the only difference between these scenarios is in starting times of the relationships. An illustration is depicted in figure 2.1.

Early models to evaluate the differences in HIV incidence between communities with only sequential monogamous partnerships and communities that have concurrent partnerships use a Markov process for formation and dissolution of relationships (Kretzschmar and Morris, 1996; Morris and Kretzschmar, 1997); as a result, therefore is no mechanism to keep track of an individuals' personal histories. This type of process generates individuals that are identical in many aspects. These models produce a cumulative degree distribution similar to a Poisson distribution (Goyal et al., 2012), which has been shown to be far from cumulative degree distributions actually observed (Handcock and Jones, 2004). Additionally, in these models the fraction of people that have simultaneous relationships approaches one as a function of time (Goyal et al., 2012). Later dynamic network models

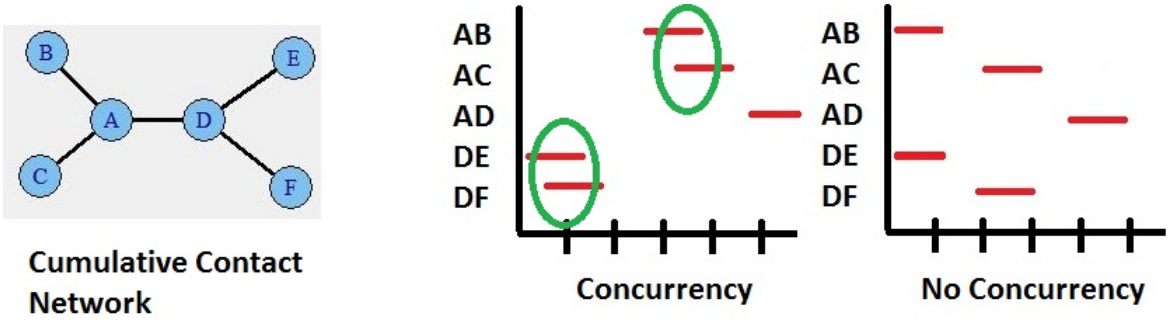

Figure 2.1: Illustration of two dynamic networks, one with concurrency and the other without concurrency. The network on the left represents the cumulative contact network. The five relationships in the network are given start and end times. The assignment for the dynamic simulation on the left has concurrency partnerships, relationships AB and AC overlap as do relationships DE and DF. The simulation on the right does not have concurrency partnerships.

developed to understand the effects of concurrency were also built under the Markov process framework and therefore have similar issues (Jeffrey W. Eaton and Garnett, 2011; Morris et al., 2009). Another drawback of the early models is that concurrency was defined such that it correlates with degree mixing, making it impossible to separate these two network properties (Jeffrey W. Eaton and Garnett, 2011; Morris and Kretzschmar, 1997). Because of these issues that arise from specifying network dynamics with momentary property measures, we instead use cumulative measures. Cumulative measures have also been shown to be better predictors of HIV transmission rates in a population (Ghani et al., 1997).

More recent studies attempted explicitly to investigate the impact of other network properties controlling for concurrency (Doherty et al., 2006; Ghani et al., 1997; Ghani and Garnett, 2000). These studies used regression methods to study the impact of network properties using simulation. Underlying assumptions of linearity and independence may not hold as many network properties exhibit sharp thresholds and are highly correlated; in addition, there exist dependencies among individuals. Therefore, the incremental impact of concurrency on incidence under different network properties cannot be well estimated using regression techniques. For a review of research on concurrency epidemic modeling see Goodreau (2011).

In our investigations three measures of concurrency were considered: The first, C_1, is the proportion of individuals with overlapping sexual partnerships at any point in the past year, referred to as "cumulative prevalence of concurrent partnerships" (Fishel et al., 2012). The second, C_2, denotes the total number of overlapping relationships. The third, C_3, is the total amount of time that relationships overlap during the past year. Since all three metrics, C_1, C_2, and C_3, represent distinct aspects of concurrency, we developed a method to fix all three measures. Mochudi, Botswana, is currently under investigation in a pilot study to evaluate the benefit of early treatment with ART to control the spread of HIV infection. The pilot study collected information on the start and end times of the responds last three relationships. Using this data we estimated values for the concurrency metrics. Appendix B describes how these three metrics of concurrency are related to alternative concurrency metrics used in other studies.

Our interest lies in investigation the impact of concurrency on the potential size of the epidemic; as a proxy, we consider the size of the largest reachable path (LRP) (Morris et al., 2009). To create reachable paths, we simulate propagation of epidemic models on the network. We start by selecting an individual to be infected and set the probability of transmission to 1. The reachable path includes the initial infected individual and the subsequent time-ordered sequence of partnerships along which transmission was possible. The time-ordered sequence is generated by sampling start and end times for relationships in the cumulative network fixing values for the concurrency metrics, C_1, C_2, and C_3. Figure 2.2 depicts an example of three dynamic simulations with identical cumulative contact networks and values for the three concurrency metrics along with the size of the associated largest reachable path for each of the simulations.

2.2 Results

To investigate the incremental effect on the length of the largest reachable path of network property f (e.g. clustering or mixing), beyond that of degree distribution and concurrency, we generated a set of network collections. The graphs in all of the collections have the same degree distribution. The set of network collections have the property that all

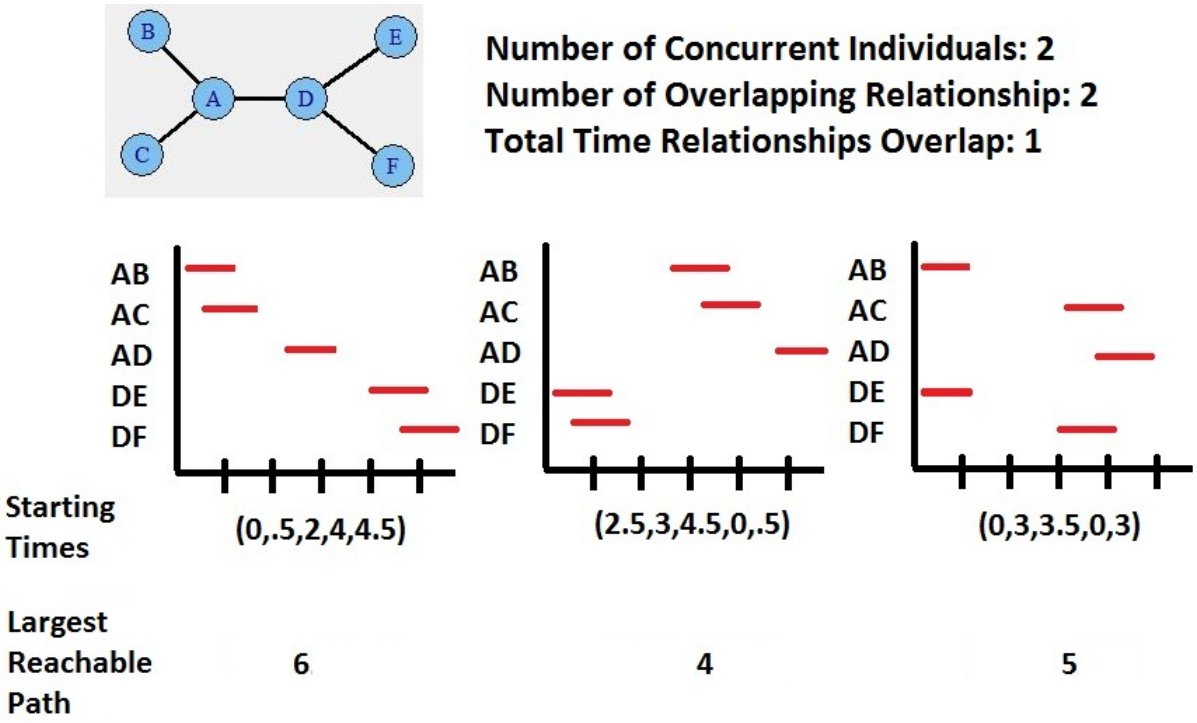

Figure 2.2: Depiction of three dynamic networks, all with identical values for concurrency measurements C_1, C_2, and C_3. The network on the top left represents the cumulative contact network. All dynamic simulations have two individuals, A and D, participating in concurrency partnerships, two pairs of relationships that overlap, and a total amount of overlap of one time unit. The start times and largest reachable path is provided for each simulation.

networks within a collection have the same value of f, but values of f differ across collections. For each network, we estimated the length of the largest reachable path by constructing dynamic networks where relationship times, randomly sampled from a uniform distribution, were consistent with given measures of concurrency (see Methods). After estimating the potential epidemic size for each specific network, we averaged these estimates across all the networks in a collection, which are composed of networks with the same value for property f. Since the degree distribution and concurrency measures are identical in all collections, the variation in average potential epidemic sizes across collections characterizes the added value of knowledge of network property f. By using the same set of network collections, and fixing the three metrics for concurrency to zero, we estimated the potential HIV incidence in populations with identical set of sexual

relationships, but for which concurrent relationships do not occur.

In this paper we study three important network properties which can be associated with partner selection behaviors: degree assortativity coefficient, clustering coefficient, and percentage of mixing between two subpopulations. Degree assortativity summaries the propensity for individuals with similar number of cumulative partners to form relationships; the exact formula can be found in Newman (2002). The clustering coefficient describes the probability of two partners of the same individual form a partnership; the coefficient is formally defined as the fraction of connected triples which are three-cycles. The percentage of mixing is defined as the percentage of all relationships in the cumulative contact network that occur between individuals in two different subpopulations of equal size.

The degree distribution used in the simulations is shown in table 2.1 and represents only individuals who were sexually active in the last year. The distribution closely resembles that observed among older adolescents in Rakai district, Uganda (Konde-Lule et al., 1997), though for simplicity we studied a one-mode network. In our simulations, the population size, n, was set at 50. From the pilot study in Mochudi, Botswana, the observed percentage of individuals with overlapping relationships was 10.04%, $C_1 = \lceil .1004 * n \rceil$; and 22.91% of individuals with at least one concurrent relationship had more than two of them, $C_2 = \lceil 1.2291 * .1004 * n \rceil$. The third metric of concurrency, C_3 was set equal to $0.64 * C_2$, which is the mean fraction of time relationships overlap (see Methods for details) multiplied by the number of relationships that are concurrent.

Table 2.1: **Distribution of Number of Partners**

	Degree Distribution				
Degree	1	2	3	4	5+
Percentage of Individuals	66	18	8	4	4

Figure 2.3 compares each network property to the length of the largest reachable path for communities with concurrency and for communities without concurrent relationships, $C_1 = C_2 = C_3 = 0$. Each black point, representing communities with concurrency, or red point, representing the communities without concurrency, was averaged over the network

collection; each collection contained five networks. The black and red lines represent the lowess curves for the scenarios with and without concurrency, respectively.

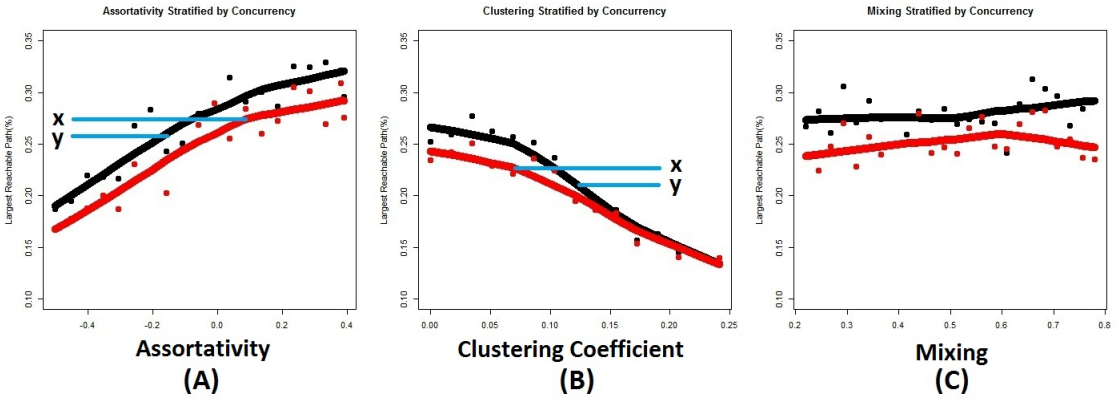

Figure 2.3: Relative importance of concurrency controlling for other network properties on the largest reachable path. The y-axis is the largest reachable path, while the x-axis is (A) degree assortativity, (B) clustering coefficient, and (C) mixing between two subpopulations.

Figures 2.3a and 2.3b demonstrate a strong relationship between potential size of the epidemic and both degree mixing and clustering; the relationship is sufficiently strong to make it possible to depict two communities, labeled X and Y, for which the community with no concurrency has a higher potential epidemic size than does the community with concurrency. Figure 2.3c shows mixing between two groups has a weak relationship with the potential size of the epidemic.

Unlike the relationship between LRP and degree mixing, the relationship between LRP and clustering as well as LRP and mixing between two communities were strongly modified by the presence of concurrent partnerships. Higher values for the clustering coefficient offset the impact of concurrency. High levels of clustering create multiple pathways for infection between individuals, thereby generating a similar the mechanism used by concurrency to increase disease spread. Since communities with high clustering already contain many multiple pathways for infection, there is diminishing marginal impact from concurrency in creating additional non-duplicate pathways. Though in this paper we consider clustering coefficient, the same logic applies for any subpopulation in a network with a high density of relationships among its members.

Figure 2.4 depicts the difference between the two curves that represent results for communities with and without concurrency in each panel of figure 2.3. As this difference is nearly a horizontal line, figure 2.4a, the predicted increase in incidence due to concurrency appears to be independent of degree assortativity. Consequently, it is possible to estimate the impact of reductions in concurrency in a community on potential epidemic size without information on degree assortativity– a difficult quantity to measure–provided that other networks properties are held fixed. By contrast, the same does not hold for either clustering or mixing between two subpopulations.

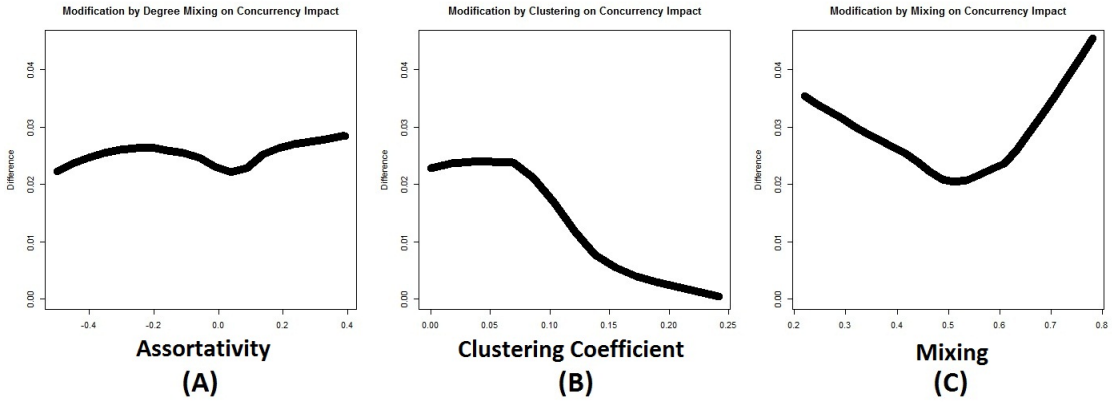

Figure 2.4: Difference between communities with concurrency and no concurrency. The y-axis is the difference. X-axis is the range of (A) degree assortativity, (B) clustering coefficient, and (C) mixing between two subpopulations

Our results demonstrate that even after controlling for cumulative ego-centric properties, degree distribution and the three metrics for concurrency, other network properties, which include degree mixing and clustering, can be very influential on LRP. In particular, high clustering is shown to completely offset the effect of concurrency. Collecting and analyzing ego-centric sexual network data can be difficult, time-consuming, and costly, and network centric data only adds to this challenge. It may not be possible to collect network centric data in settings of interest, though studies have succeeded in doing this (Helleringer and Kohler, 2007). In any case, it is critical to understand the way in which information regarding network centric data can contribute to an understanding of the HIV epidemic as well as the limitations of conclusions drawn in the absence of such information.

2.3 Discussion

We applied our methods to simplified versions of complex interacting systems that depend both on population behavior and on biological processes. We considered only reachable paths and did not attempt to include the biological characteristic of HIV transmission that are relevant in actual epidemics, such as the increased risk of transmission during the acute HIV infection or decreases in risk of transmission due to coital dilution (Sawers et al., 2011). However, these details are not necessary to make our fundamental point: network properties that are not estimable from ego-centric data are necessary to make conclusions about the impact of concurrency on epidemic characteristics and on the efficacy of measures to reduce concurrency. To make this point, we developed a novel method to sample dynamic networks constrained by particular network properties. This method allowed us to estimate a proxy for epidemic size in communities with different measures of concurrency. While controlling for the cumulative contact network, we considered two extreme scenarios, high levels of concurrency and no concurrency (recognizing the impossibility of achieving the latter in an actual community), to depict the strong impact of other network properties relative to that of concurrency. The properties studied in the paper were selected because they represent behaviors that reflect plausible partner selection criteria.

The strong influence of other network properties provides a possible explanation for the "mysterious" (Epstein and Stanton, 2010) phenomenon that communities with higher levels of polygamous relationships have lower HIV rates than do others, despite the fact that those actually participating in polygamous relationships have higher HIV rates than do other community members. The cumulative contact network of communities with higher levels of polygyny will be tend to be more dissassorative, since polygyny requires high degree individuals paired with individuals with only one partner. Due to the strong relationship between HIV incidence and degree assortativity, as depicted in figure 2.3a, dissassorative communities will have lower HIV incidence compared to similar communities which are more assortative. Nonetheless, concurrency due to polygynous relationships is not "benign" (Reniers and Watkins, 2010), but increases the risk of HIV among those mem-

bers of the polygamist family. This can be understood by viewing the polygamist family as forming a community in themselves; as shown in figure 2.3a, concurrency always increases the potential size of the HIV epidemic in a community. In summary, changes in the unobserved network properties–in particular degree assortativity– can explain communities with higher levels of polygyny may have lower overall HIV rates despite higher HIV rates within the members of polygamist relationships. Our results compliment ideas presented in Kretzschmar et al. (2010).

The results highlight concern about the impact concurrency reduction programs. Changes in concurrency patterns can, and probably will, cause changes in the cumulative network which in turn will modify network measures. Therefore, the impact of the concurrency reduction on HIV incidence is not really predictable without measuring its impact on other network features; in fact it is even possible to increase the potential epidemic size after reduction of concurrency, as indicated by two hypothetical communities, labeled X and Y, in figure 2.3a. An example of such a scenario is one in which a former polygamist maintains the same high number of cumulative sexual partners, but is not able to find partners with only a single cumulative partner as was the case for the concurrent polygamous partnerships.

Interventions to reduce concurrency will likely reduce the spread of HIV (except in extreme clustering), provided that the cumulative contact network does not change. However, the reduction in HIV incidence observed in this paper, where we exactly controlled for the cumulative network and three different aspects of concurrency, was not high as that investigated in other models (Morris et al., 2009).

2.4 Materials and Methods

The following section assumes a cumulative contact network G, and provides details on how to calculate the expected size of the largest reachable path vector if we sample start and end times for each relationships randomly from a uniform distribution such that each set of start and end times are consistent with the three values of concurrency, C_1, C_2 and C_3. Figure 2.2 depicts a simple example.

2.4.1 Graph Theory Terminology

Some definitions from graph theory will be useful to relate measures of concurrency to network literature. Let the cumulative contact graph be represented by $G = (V, E)$, where V is the set of individuals in the population and E is the set of relations between individuals, $E \subset V \times V$. In this paper, we define G to represent all the relationships that occurred during a fixed period of time of length T.

Definition 1: Contact Graph - A contact graph is a representation of relationships in a fixed population. The vertices in the graph represent individuals in the population and an edge exists if there exists a relationship between the two individuals.

Definition 2: Line Graph - A line graph is a representation of the relationships between the edges in a contact graph. Let $L(G)$ be the line graph of a graph G. The vertices of $L(G)$ are taken as the edges of G, and two vertices of $L(G)$ are adjacent whenever the corresponding edges of G are adjacent (Harary, 1969).

$L(G)$ will be useful in dynamic network simulations. For example, to simulation networks with fixed C_2, we can sample a fixed number of edges. E_o, to be labeled "overlapping" in time. The remaining edges in $L(G)$, E_n, are labeled as "non-overlapping". Not every sample of labeled edges has a valid realization, i.e. pairs of real values representing the start and end time of each relationship such that the labels are preserved. To assess requirements for a labeled $L(G)$ to have a valid realization we introduce the following terminology.

Definition 3: Interval Graph - A graph, $G = (V, E)$, is called an interval graph if there $\exists$ a set of intervals $\{I_j\}$ such that $I_l \cap I_k \neq 0$ if and only if $(l, k) \in E$ (Fishburn, 1985).

Assuming $L(G)$ is a complete graph (i.e. there exists an edge between each pairs of vertices), it can be shown that $L(G)$ has a valid realization if and only if the graph

containing only edges labeled ”overlapping" is an interval graph. However, in sexual networks, $L(G)$ will be sparse, and therefore not complete. The edges absent from $L(G)$, $(E_o \cup E_n)^c$, have no individual in common; therefore, concurrency measures are invariant to whether or not these relationships overlap. In order to handle sparse $L(G)$, the following terminology is introduced (Golumbic et al., 1995).

Definition 4: Sandwich Graph - Let E_1 and E_2 be two disjoint sets of edges defined on the same vertex set V. A graph $G = (V, E)$ with $E_1 \subseteq E \subseteq E_1 \cup E_2$ is called a sandwich graph for (E_1, E_2)

Definition 5: Interval Graph Sandwich Problem - Given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ defined on the same set of vertices but with disjoint edges-sets E_1 and E_2, does there exist a sandwich graph for (E_1, E_2) that is an interval graph? (Golumbic et al., 1995)

$L(G)$ has a realization if and only if there is a solution to the interval graph sandwich problem defined by $G_1 = (V, E_1 = E_o)$ and $G_2 = (V, E_2 = (E_o \cup E_n)^c)$, because E_o represents relationships constrained to overlap while $(E_o \cup E_n)^c$ represents all pairs of relationships that are inconsequential to concurrency measures with regards to whether they overlap or not. Ascertainment of whether or not a graph is an interval graph is computationally efficient, but the validation of a solution to the interval graph sandwich problem can be computationally intense (Golumbic et al., 1995). Due to the sparsity of $L(G)$ it may be possible to efficiently generate a solution to the interval graph sandwich problem satisfying constraints imposed by the labeled $L(G)$.

In addition to labels, non-overlapping edges, $(E_{ij}, E_{ik}) \in E_n$, are also associated with a direction, $d_{(E_{ij},E_{ik})} \in \{<, >\}$, indicating which relationship, E_{ij} or E_{ik}, occurs first in time order. Let $s(E_{ij})$ and $e(E_{ij})$ denote the start and end times of the relationship between individuals i and j. If $d_{(E_{ij},E_{ik})} = '<'$ then $s(E_{ij}) < s(E_{ik})$, otherwise if $d_{(E_{ij},E_{ik})} = '>'$ then $s(E_{ij}) > s(E_{ik})$.

Relating Labeled Line Graph and Associated Directions to $\mathbb{R}^{|E|}$

The largest reachable path given labels and directions attributed to relationships is not affected by relationship duration; this statement can be considered a corollary to Theorem 1. Therefore, for simplicity of notation, we assume all relationships have the same length; without loss of generality this length can be a single unit. Relationship duration can affect the mean largest reachable path (see below); nonetheless, all of the theory presented below holds for any distribution of relationship durations.

The labeled graph and associated directions has a corresponding convex subspace of $\mathbb{R}^{|E|}$, in particular the subspace is a convex polytope, defined by the linear constraints described below. Since all relationships have unit length $e(E_{ij}) = s(E_{ij}) + 1$. Each edge, (E_{ij}, E_{ik}), is associated with one of three designations: Overlapping, Non-overlapping with $d_{(E_{ij},E_{ik})} = '<'$, and Non-overlapping with $d_{(E_{ij},E_{ik})} = '>'$. If an edge in $L(g)$, (E_{ij}, E_{ik}), is labeled overlapping then $s(E_{ij}) + 1 > e(E_{ik})$ and $s(E_{ik}) + 1 > e(E_{ij})$. Otherwise, the edge (E_{ij}, E_{ik}) is labeled non-overlapping, requiring $|s(E_{ij}) - s(E_{ik})| > 1$. The direction of non-overlapping edges distinguishes between the two possibilities, $s(E_{ij}) + 1 < s(E_{ik})$ or $s(E_{ik}) + 1 < s(E_{ij})$.

Let $V^p(A)$ be the set of nodes that are reachable from node A using the set of relationship start times defined by point $p \in \mathbb{R}^{|E|}$. Therefore, the size of the largest reachable path for time set p is $LRP(p) = Max\{|V^p(A)| : A \in V\}$.

2.4.2 Algorithm to Estimate Mean Size of the Largest Reachable Path Given G and Fixed Concurrency Values

Outline

The algorithm to calculate the mean length of the largest reachable path of G is based on sampling subspaces of $\mathbb{R}^{|E|}$, where the coordinates for each point in the subspace represent start times for the set of relationships in G. Let $S \subset \mathbb{R}^{|E|}$ be the subspace such that each point, $s \in S$, is consistent with the three values of concurrency, C_1, C_2 and C_3. Let $LRP(S)$ denote the mean length of the largest reachable path for the subspace S. $LRP(S)$ is equal to the following,

$$S = LRP(S) = \frac{1}{Vol(S)} \int_S LRP(s)\mathrm{d}s \tag{2.1}$$

where $Vol(S)$ denotes the volume of the subspace S. Because, sampling points directly from S is a complex process, we partition S into subspaces.

Let $W \subset \mathbb{R}^{|E|}$ be the set of points that fix C_1 and C_2 only; thus, $S \subset W$. Let $\{H\}_{\alpha \in A}$ be a partition of W such that each element, H_α, is uniquely defined by the subspace which corresponds to a labeled line graph and associated directions. We further partition each H_α into subspaces, $\{K\}_{\beta \in B_\alpha}$, such that the reachable paths for all points in K_β are identical, i.e $V^{p_1}(A) = V^{p_2}(A) \, \forall \, A \in V(G)$ if p_1 and p_2 are in the same partition K_β. Theorem 1 proves that the partition of H_α, $\{K_\beta\}_{\beta \in B_\alpha}$, corresponds simply to including additional labeled edges and associated directions beyond those already attributed to $L(G)$. Define a partition $\{X\}_{\beta \in B}$ of S, where $X_\beta = K_\beta \cap S$. Therefore,

$$S = \bigcup_{\alpha \in A} \left(\bigcup_{\beta \in B_\alpha} X_\beta \right) \tag{2.2}$$

Because $S \subset \bigcup_{\alpha \in A} H_\alpha$,

$$S = \left(\bigcup_{\alpha \in A} H_\alpha \right) \bigcap \left[\bigcup_{\alpha \in A} \left(\bigcup_{\beta \in B_\alpha} X_\beta \right) \right] \tag{2.3}$$

Substituting the partition $\{K_\beta\}_{\beta \in B_\alpha}$ for H_α,

$$S = \left[\bigcup_{\alpha \in A} \left(\bigcup_{\beta \in B_\alpha} K_\beta \right) \right] \bigcap \left[\bigcup_{\alpha \in A} \left(\bigcup_{\beta \in B} X_\beta \right) \right] \tag{2.4}$$

Using the distributive law, which applies because $|A|$ and $|B_\alpha| \, \forall \, \alpha \in A$ are finite, we get the following,

$$S = \bigcup_{\alpha \in A} \left[\bigcup_{\beta \in B_\alpha} \left(K_\beta \cap X_\beta \right) \right]. \tag{2.5}$$

Substituting (2.5) into equation (2.1),

$$LRP(S) = \frac{1}{Vol(S)} \int_{s \in \cup_{\alpha \in A} [\cup_{\beta \in B_\alpha} (K_\beta \cap X_\beta)]} LRP(s)\mathrm{d}s. \tag{2.6}$$

Elements of the set $\{H_\alpha\}_{\alpha \in A}$ form a partition of S, therefore,

$$LRP(S) = \frac{1}{Vol(S)} \sum_{\alpha \in A} \int_{s \in \cup_{\beta \in B_\alpha}(K_\beta \cap X_\beta)} LRP(s)\mathrm{d}s \tag{2.7}$$

Similarly, elements of the set $\{K_\beta\}_{\beta \in B_\alpha}$ define a partition of B_α, so,

$$LRP(S) = \frac{1}{Vol(S)} \sum_{\alpha \in A} \sum_{\beta \in B_\alpha} \int_{s \in K_\beta \cap X_\beta} LRP(s)\mathrm{d}s \tag{2.8}$$

Because the size of the largest reachable path is constant in K_β,

$$LRP(S) = \frac{1}{Vol(S)} \sum_{\alpha \in A} \sum_{\beta \in B_\alpha} Vol(K_\beta \cap X_\beta) * LRP(K_\beta) \tag{2.9}$$

$$= \sum_{\alpha \in A} \sum_{\beta \in B_\alpha} \frac{Vol(K_\beta \cap X_\beta)}{Vol(S)} * LRP(K_\beta) \tag{2.10}$$

Since we are interested in uniformly sampling points (i.e. sets of start times for the dynamic network simulation) that are consistent with the three metrics of concurrency, the probability of sampling a subspace is proportional to the volume. Therefore,

$$LRP(S) = \sum_{\alpha \in A} \sum_{\beta \in B_\alpha} P(K_\beta \cap X_\beta) * LRP(K_\beta) \tag{2.11}$$

$$\propto \sum_{\alpha \in A} \sum_{\beta \in B_\alpha} P(K_\beta) * P(X_\beta | K_\beta) * LRP(K_\beta) \tag{2.12}$$

where $P(K_\beta)$ is defined proportionally to the volume of K_β.

Algorithm 1, presented below, provides an outline of the procedure to estimate the mean reachable path given G and concurrency values, C_1, C_2 and C_3. To estimate LRP(S) using equation (2.12), subspaces of the form K_β need to be sampled proportional to $P(K_\beta) * P(X_\beta|K_\beta)$. Algorithm 1 (steps 2-4) starts by generating a sample, H_α, which is constructed by assigning labels and directions to edges of $L(G)$. Step 5 creates the partition, $\{K_\beta\}_{\beta \in B_\alpha}$, of the space H_α, along with sampling an element, K_β, from the partition. Steps 2-5 generate a sample of K_β under a probability distribution, Q, which is different from a probability distribution defined by sampling K_β proportional to $P(K_\beta) * P(X_\beta|K_\beta)$. Therefore, importance

weights are needed to appropriately adjust for the procedure used in steps 2-5 to sample subspaces; the re-weighing of the sample is performed in step 7 of algorithm 1. The weights are based on the probability of generating a sample for K_β ($Q(K_\beta)$), the probability of K_β ($P(K_\beta)$), and the density of points in K_β that fix C_3 ($P(X_\beta|K_\beta)$). As mentioned above, mean LRP is affected by relationship duration as the probabilities of sampling labeled line graphs and associated directions may depend on relationship durations. But our conclusions regarding the fundamental effects of network properties on LRP are unaffected by relationship durations.

```
Algorithm 1: Estimate mean length of largest reachable path
Input: Graph g, C1, C2, and C3
Output: Expected reachable path with fixed C1, C2, and C3
1. Construct L(g)
repeat {
    2. Label edges in L(g) as "overlapping" fixing C1 and C2
    3. Verify labeled L(g) has a valid realization, if not repeat step 2
    4. Generate H by assigning directions to "non-overlapping" edges
    5. Generate K by sampling a partition of H
    6. Estimate the proportion of K that is consistent with C3
}
7. Compute importance weights of sampled subspaces.
```

In several steps of Algorithm 1, we need to sample points from a convex polytope, a convex polygon in higher dimensions. For completeness we present a previously published method to sample points uniformly from a polytope (Smith, 1984).

```
Algorithm 2: Sample Uniformly from a Convex Polytope, P
Input: A polytope either as halfspace or points
Output: Set of points, {p1,p2,..pn}, sampled uniformly from P
1. Let p=p0 be any point in P, and i = 1
repeat {
```

```
    2. Construct a random line, l, through the point p
    3. Uniformly select a point, pi, on the line l
    4. Increment i = i + 1
    5. Let p = pi
}
```

Below we describe in detail each of the steps in Algorithm 1.

Step 1: Construction of L(G)

The construction of L(G) is a straight forward procedure. In our algorithm all edges in L(G) will be characterized by a label, overlapping or non-overlapping. For those edges labeled as non-overlapping, the edge is associated with a direction. As previously stated, $L(G)$ provides information on the relationship between edges in G. The labeled graph with directions has a corresponding subspace of $\mathbb{R}^{|E|}$, in particular a convex polytope, by using the linear constraints described above. Therefore, we denote $L(g)$ with the associated labels and directions as both a line graph and the corresponding subspace.

Step 2: Select Individuals and Relationships

Select C_1 individuals, N_c, from $G = (V, E)$ and C_2 edges, E_o, from $L(G)$ with the following two criteria.

- For any edge $(E_{ij}, E_{ik}) \in E_o$, i must exist in N_c
- For any individual $i \in N_c$ there exists an edge $(E_{ij}, E_{ik}) \in E_c$

Label the selected edges, E_o, as ”overlapping" and the unselected edges as ”non-overlapping".

Step 3: Verify Labeled $L(G)$ has a Valid Realization

There may not be a realization of the labeled line graph $L(G)$. Since in our example the proportion of concurrent partnerships is small compared to the total number of relationships, we modify a previously published validation method (Pe’er and Shamir, 1995) to make

the computation feasible. Validity can be checked separately on each connected components of a graph that includes only edges labeled as "overlapping". Let $SL(G) \subset L(G)$ be a subgraph of $L(G)$ defined by including only edges in $L(G)$ labeled as overlapping. Assigning a distinct block of time for each component in $SL(G)$, makes clear that if validity holds for each components of $SL(G)$, then validity holds for $L(G)$. Since $C_2 \ll |E|, SL(G)$ is extremely sparse and therefore both the average degree and the expected second-order average degree are typically less than one. Hence, the components of $SL(G)$ are small (Chung and Lu, 2002) and it can be easily verified whether a valid realization exists.

Step 4: Assign Direction to Non-Overlapping Edges

Each non-overlapping edge is assigned a direction. The direction on an edge identifies which of the relationships represented by the endpoints precedes the other relationship in time. First, direction is established for "non-overlapping" edges in $L(G)$, that are contained completely in a connected component of $SL(G)$ as outlined in Pe'er and Shamir (1995). The remaining "non-overlapping" edges of $L(G)$ without a direction assignment join distinct connected components of $SL(G)$. The assignment of direction of the remaining edges can be done by starting with an initial component of $SL(G)$ and sequentially adding one additional component. At each step of adding a component, "non-overlapping" edges are assigned a valid direction. Using this procedure it is possible to guarantee a valid label and direction for $L(G)$, as long as each component of $SL(G)$ has a valid realization. We will refer to $L(G)$, as constructed at the end of step 4, and the corresponding subspace as H.

Step 5: Add edges to H in order to have consistent largest reachable path

The length of the largest reachable path and the value of C_3 can be evaluated for all points in H, each of which represent the starting times of the set of relationships. Because the joint density distribution of the starting times, C_3, and largest reachable path is not smooth, it is difficult to estimate this distribution. For this reason we partition H in subspaces in which all points have the same reachable paths. From the partition of H we select one subspace K. Restricting consideration to K also decrease computation time as the largest reachable path does not need to be evaluated for each point in the K. Theorem

1 describes the necessary additional labeled edges required to find the subspace K.

Theorem 2.1: The following information characterizing relationships in a given cumulative contact network $G = (V, E)$ is necessary and sufficient to construct a subspace K such that $V^{p_1}(A) = V^{p_2}(A)\ \forall\ p_1, p_2 \in K$ and $\forall\ A \in V$.

1. Listing of overlapping edges
2. Direction of non-overlapping edges
3. Labels and directions of edges completely contained in connected components of $L(G)$ using only overlapping edges.

The proof can be found in Appendix C.
To satisfy these constraints, we sample a realization of H using Algorithm 2 and add in the necessary labeled edges using the realization. Let K denote both the graph $L(G)$ with additional edges to H and the corresponding subspace.

Step 6: Estimate the proportion of K that is consistent with C_3

The polytope K is a region representing start times that is consistent with two measures of concurrency, C_1 and C_2. In addition, all points in K have the same largest reachable path. However, the region does not fix the currency value C_3, the total amount of time relationships overlap. Since we are interested in estimating the expected size of the largest reachable path of G, it is sufficient to estimate the proportion of K that is consistent with C_3, $P(X|K)$. Algorithm 2 provides a procedure to sample points from K. Once a set of points is sampled from K, it is straight forward to calculate C_3. Since the joint distribution between C_3 and the starting times of relationships is smooth, is it possible to use kernel density estimation procedures Rosenblatt (1956) to approximate the proportion of K that is consistent with C_3.

7: Compute importance weights of sampled subspaces

To evaluate equation (2.12), subspaces K need to sampled proportional to $P(K) * P(X|K)$; but steps 2-5 samples of K were generated under a different distribution. Therefore, we use importance weights to adjust for the discrepancy between the two probability distributions. Let Q represent the probability distribution for sampling subspaces as described in steps 2-5. To generate a sample K, we first generate an H (steps 2-4). In step 5, we sample K proportional to volume of K in H, which is $P(K|H)$. Therefore, $Q(K) = Q(H) * P(K|H)$

The formula to calculate the estimated expected reachable path vector for a graph G uses the following weights for each sampled subspace, $K_1, .., K_m$, from steps 2-6.

$$w_i = \frac{h(K_i)/g(K_i)}{\sum_{j=1}^{m} h(K_j)/g(K_j)}$$

where,

$$h(K_i) = P(K_i) * P(X_i|K_i) \tag{2.13}$$

and,

$$g(K_i) = Q(K_i) = Q(H_i) * P(K_i|H_i) \tag{2.14}$$

Since step 6 calculated $P(X_i|K_i)$, we only need to calculate $P(K_i)$, $Q(H_i)$, and $P(K_i|H_i)$.

Under the uniform distribution, $P(K_i) \propto Vol(K_i)$. To calculate the exact volume of a polytope is very difficult (Bárány and Füredi, 1987); an active research area in computational geometry is approximating this volume (Dyer et al., 1991). One approach is to approximate the volume by generating spaces $K_i = R_p \subseteq \cdots \subseteq R_1$, where the volume of R_1 is easy to calculate. Once the nested subspaces are defined, the volume of K_i can be rewritten as the following:

$$Vol(K) = Vol(R_1) * \Pi_{j=1}^{p-1} \frac{Vol(R_{j+1})}{Vol(R_j)}$$

The ratio of $\frac{Vol(Rj+1)}{Vol(R_j)}$ for $j \in 1, \cdots, p$ can be estimated using the proportion of sampled points in R_j that are also contained in R_{j+1}.

In our simulation R_1 is the space defined by distinct paths in $L(G)$, $p_1, \cdots, p_m$, that cover all the vertices and use only edges labeled as "non-overlapping" in H_i. Thus,

$$Vol(R_1) = \Pi_{j=1}^{m} \frac{1}{|p_j|} * (-(|p_j| - 1) * l + T)^{|p_j|},$$

where l is the length of a relationship and T is the time period of interest for the study (in our case $l = 1$ month and $T = 12$ months). R_{j+1} is defined as R_j plus one additional edge from K. Since R_1 is a subset of edges contained in H_i, $\exists$ a j such that $R_j = H_i$. So, it is straightforward to calculate $P(K|H) = \frac{Vol(K)}{Vol(H)}$. Finally, $Q(H_i)$ can be computed using a brute force method.

Steps 2-5 of Algorithm 1 provide a method to sample a subspace, K, which satisfy concurrency metrics C_1 and C_2. Steps 6-7 adjust the probability of sampling the subspace in order to be proportional to $P(K) * P(X|K)$, which is necessary to evaluate equation (12). Therefore, we can approximate the LRP for the subspace that fixes C_1, C_2 and C_3.

The Importance of Modeling Degree Mixing in HIV Network Simulation Models

Ravi Goyal, Joseph Blitzstein, and Victor DeGruttola

Department of Biostatistics

Harvard School of Public Health

3.1 Introduction

Features of the networks along which disease propagates may impact the efficacy of interventions intended to prevent or control communicable infectious disease epidemics (Wylie and Jolly, 2001). Therefore, network-based simulation studies of the effects of prevention programs may be useful for designing such interventions, as well as for making policy decisions about their deployment. Hence, interest lies not only in estimating social or sexual network properties, but also in forming a collection of networks, where selection into the collection is based upon the probability of the network being the true realization for the unknown network of interest.

Collection of information regarding sexual networks is challenging; as a consequence, most studies consider only ego-centric network properties–degree distribution and non-degree mixing patterns that can be estimated from ego-centric survey data. By contrast, estimation of degree mixing patterns typically requires a form of link tracing making collection of such information more challenging. Nonetheless, degree mixing has been shown to be of particular importance in a variety of research areas including properties of the Internet (Doyle et al., 2005; Vázquez et al., 2002) and of biological interactions (Maslov and Sneppen, 2002). In disease transmission models, Newman (2002) concluded that degree assortative networks disseminate disease more easily and are more robust to removal of their highest degree nodes compared to disassortative networks. This insight may have important implications for HIV prevention programs, in particular regarding the efficacy of treatment of HIV-infected individuals as prevention. In chapter 2, it was shown that degree mixing can have a large impact on transmission even after degree distribution and concurrency are accounted for in the modeling.

Despite the challenges in collecting network centric data in settings of sexual disease transmission, some studies have succeeded in doing so (Helleringer and Kohler, 2007). Whether or not such information can be collected, however, it is important to understand the way in which information regarding degree mixing can contribute to knowledge of HIV epidemic dynamics as well as the consequences for this lack of information on the precision of inferences that can be drawn from network simulation studies. The peaked

distribution of many network properties, including degree mixing, can cause a parameter that is not explicitly modeled to be implicitly assigned an essentially fixed value. Figure 3.1 depicts an example of the distribution of the degree assortativity coefficient (Newman, 2002), a summary metric for the degree mixing matrix that takes on values between -1 and 1, which is sharply peaked when the degree distribution is fixed. This assigned value for degree mixing is valid when the partner selection process is independent of an individual's previous number of partners, conditioned on observed covariates. Since partner selection is a complex process, such a requirement may be hard to justify.

In order to accommodate network properties with peaked distributions, we develop a method to model influential parameters with varying degrees of uncertainty. The method considers two scenarios: one where information is available to characterize degree mixing and one where it is not. Results from the latter demonstrate that if degree mixing parameters are highly uncertain, the ability to make inferences regarding outcomes of epidemiological models is adversely affected. Therefore, reliable inference from network models requires new methods to bound values for unknown parameters rather implicitly set these parameters at a fixed value; such methods are described below.

Section 3.3 develops a method to construct a collection of bipartite networks based on a probability function for selecting networks into the collection. The presented method extends the framework presented in chapter 1 to explicitly model degree mixing, assuming a fixed estimate for the degree distribution. Chapter 1 used parameter estimates and their estimated variance to construct one-mode networks; parameters included density, degree distribution and mixing patterns. Appendix D extends these methods to identical settings but for bipartite networks.

Section 3.4 provides a method for estimating degree mixing from a sample given an estimate of the degree distribution, and section 3.5 compares a pair of collections of bipartite networks with a fixed degree sequence–one based on sampled mean and variance estimates of degree mixing patterns and the other, on a uniform distribution for degree mixing patterns. Section 3.6 provides a discussion.

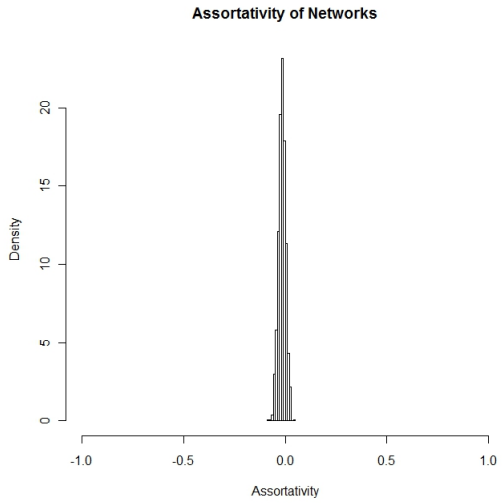

Figure 3.1: Distribution of the degree assortativity coefficient. The histogram is constructed by uniformly sampling networks with the same degree distribution as collected from sexual network study in Likoma Island.

3.2 Network Construction

Our discussion of network construction expands upon the framework presented in chapter 1, which provided a method to model degree distribution and both degree and non-degree mixing patterns for one-mode networks. The framework allowed different measures of uncertainty for each parameter; our extension also include this feature. Our goal in this paper is investigation of the usefulness of making use of estimated degree mixing parameters compared to simulating a uniform distribution of degree mixing values. Therefore, our models fix degree distribution and include only parameters for degree mixing. We will also focus on bipartite graphs, in which vertices are designated either male (m) or female (f). Appendix D expands the method for density, degree distribution, and mixing patterns for bipartite networks without the requirement of the degree distribution being fixed.

To describe the method for constructing network collections requires defining terminology and notation. Let vector $D^t(g)$ denote the degree distribution for nodes of type $t \in \{m, f\}$ of a graph g, where the i^{th} entry of $D^t(g)$, $D^t_i(g)$, is the number of type t nodes with degree $i-1$. Let $d^t(g)$ represent the degree sequence for type t nodes in network g, where the i^{th} entry, $d^t_i(g)$, is the degree of i^{th} type t node. Let $MM(g)$ be a matrix representing the mixing

pattern of graph g. The entry $MM_{k,l}(g)$ is the number of edges from a male with covariate pattern k to a female with covariate pattern l. We use the notation $DMM(g)$ to represent degree mixing matrices, where entry $DMM_{i,j}(g)$ is the number of edges from a male of degree i to a female of degree j.

The network collection is a subset of networks from the space, $\mathscr{G}$, of bipartite graphs with n nodes and fixed degree distributions of D^m and D^f. To construct such a collection, we begin by partitioning $\mathscr{G}$ into congruence classes, such that each network in the congruence class has the same values for degree mixing matrix. Let C_g represent the congruence class containing network g. Therefore, networks g and h will reside in the same congruence class if and only if $DMM(g) = DMM(h)$. Each network in $\mathscr{G}$ will be assigned a probability, $P_{\mathscr{G}}(g)$, of being selected into the collection. $P_{\mathscr{G}}(g)$ is based solely on the congruence class of the network g.

By partitioning $\mathscr{G}$ into congruence classes and defining a probability mass function for the probability of sampling a network from a congruence class we can control the probability of sampling a network with particular values for network properties. As seen in figure 3.1, some congruence classes have vastly more networks than do others, this approach guards against over or under representing networks with particular properties due to the size of the congruence class, thus ensuring that the collection of networks is consistent with the collected data. Defining the probability of sampling networks from a congruence class allows for construction of networks that reflect both the estimated mean and uncertainty associated with the estimate– the only information available to the investigator– without requiring consideration of the complex topology of the underlying space of graphs of size n with a fixed degree distribution.

A Markov chain Monte Carlo (MCMC) procedure is the basis for generating a collection of networks, $\{g_1, \cdots, g_t\}$ that satisfy the probability distribution assigned to the congruence classes. Ideally, to construct our collection, $\{g_1, \cdots, g_t\}$, we would sample, with replacement, t congruence classes $\{C_1, \cdots, C_t\}$, based on the probability distribution on the classes. For each congruence class, C_i where $i \in \{1, \cdots, t\}$, we would draw a network, g_i, such that $g_i \in C_i$. Since this procedure presents computational difficulties, we implement a Markov chain using the Metropolis-Hastings algorithm to generate the networks. For a review on

MCMC methods see Robert and Casella (2004). In order to implement the Metropolis-Hastings algorithm, four aspects have to be specified: target function, proposal function, acceptance probability, and initial starting element. Many authors have described construction of an initial starting element (Blitzstein and Diaconis, 2010), so we discuss only the first three aspects below.

3.2.1 Target Function

The target function is the desired stationary distribution for the Markov chain. In our setting, the network g has a probability mass equal to the probability of the congruence class C_g divided by the number of networks in C_g, $|C_g|$, thereby ensuring that each network in C_g has the exact same probability:

$$P_{\mathscr{G}}(g) \propto \left(\frac{1}{|C_g|}\right) * P_C(C_g). \tag{3.1}$$

Due to the constraints imposed on bipartite networks given a degree distribution not all values of degree mixing matrices correspond to valid networks. Theorems 1 below gives criteria for determining whether a degree mixing matrix D is graphical for either a simple undirected bipartite network given a degree distribution.

Theorem 3.1: *An matrix, DMM, is graphical by a bipartite undirected network if and only if the following four conditions are met given degree distributions D^mandD^f:*

1. $D_i^f := (\sum_j DMM_{(i,j)})/i \in \mathbb{Z}^+ \forall i$
2. $D_j^m := (\sum_i DMM_{(i,j)})/j \in \mathbb{Z}^+ \forall j$
3. $DMM_{(i,j)} \leq D_i^f * D_j^m$
4. $DMM_{(i,j)} \geq 0$

Refer to the Appendix E for the proof of Theorem 3.1.

3.2.2 Proposal Function

The algorithm generates network g_{t+1} based only on the previous network g_t by nominating a potential network given g_t. A common algorithms used to propose a network from a given network in the space of fixed degree sequence is edge switching. The algorithm selects two edges at random, (a, b) and (c, d), from g_t. If the edges (a, d) and (c, b) do not create multiple edges or self loops, the network, p_{t+1}, which is created by replacing edges (a, b) and (c, d) with (a, d) and (c, b) is proposed. Otherwise the proposed network, p_{t+1}, is just g_t. To ensure that the edge switching procedure produces a bipartite network, nodes a and c must be of the same type, similarly, for b and d. The algorithm produces an irreducible Markov chain among all graphs with fixed degree sequence. The chain also has equal forward and backward probabilities.

3.2.3 Acceptance Probability

Once a proposal network, p_{t+1}, is generated the Metropolis-Hastings algorithm will either accept, $g_{t+1} = p_{t+1}$ or reject, $g_{t+1} = g_t$, the proposal. The Metropolis-Hastings acceptance probability is the following which was derived in [Paper 1]:

$$P(\text{Accept } gp_{t+1}|g_t) = \frac{f(C_{gp_{t+1}}, C_{g_t})}{f(C_{g_t}, C_{gp_{t+1}})} * \frac{P_C(C_{gp_{t+1}})}{P_C(C_{g_t})} \tag{3.2}$$

where $f(C_g, C_h)$ as the average number of elements in C_h that are valid proposals from an element $g \in C_g$. The value of $f(C_g, C_h)$ can be calculated from the degree mixing matrices, $DMM(g)$ and $DMM(h)$, associated with C_g and C_h. Since we only are interesting in the ratio of $f(C_g, C_h)$ and $f(C_h, C_g)$, we will assume that $C_g \neq C_h$, otherwise the ratio will be one. Given that $C_g \neq C_h$, $f(C_g, C_h), f(C_h, C_g) > 0$ only if$DMM(g)$ and $DMM(h)$ have exactly four different entries,$(j, i), (l, k), (j, k)$ and (l, i), such that the following relationships hold:

$$DMM_{(j,i)}(g) = DMM_{(j,i)}(h) - 1$$
$$DMM_{(l,k)}(g) = DMM_{(l,k)}(g) - 1$$
$$DMM_{(j,k)}(g) = DMM_{(j,k)}(g) + 1$$
$$DMM_{(l,i)}(g) = DMM_{(l,i)}(g) + 1.$$

Proposition 3.1: *Given* $C_g \neq C_h$, $f(C_g, C_h) \approx DMM_{(j,k)}(g) * DMM_{(l,i)}(g) * (1 - P_1 - P_2 + P_1 * P_2)$, *where* $P_1 = \frac{(i-1)*(j-1)*DMM_{(j,i)}(g)}{(i*D_i^f-1)*(j*D_j^m-1)}$ *and* $P_2 = \frac{(k-1)*(l-1)*DMM_{(k,l)}(g)}{(k*D_k^f-1)*(l*D_l^m-1)}$.

The proof of proposition 3.1 is located in Appendix F. Using proposition 3.1, we are able to calculate the acceptance probability for each proposal graph in our MCMC.

3.3 Estimation

In order to compare networks generated with and without a sampled estimate of degree mixing, we propose the follow sampling and estimation method. In certain settings an estimate of degree mixing maybe possible. One possibility is in testing centers where HIV positive individuals are asked to encourage current and former partners to come to the testing center. Though this will surely provide a biased sample, we will ignore this issue and propose the following simplified version of the sampling scheme. The estimation procedure describe in this section will work for any sampling design provided that the probability of sampling an edge with endpoint degrees i and j, $\pi_{(i,j)}$, depends only on i and j.

- Select n individuals at random, $\{v_1, \cdots, v_n\}$, from the network with replacement and observe the degree, e_{k_1} for each node, v_k.
- For each sampled node, select an edge at random, to be traced to the other endpoint and observe that endpoint's degree, e_{k_2}

Instead of estimating DMM, the degree mixing matrix, we will estimate $\mu = DMM / \sum_j \sum_i DMM_{(i,j)}$, where $\mu_{(i,j)}$ represents the percentage of edges between a female node of degree i with a male node of degree j. We begin with some notation: Let e_k, for $k = \{1, \cdots, n\}$, denote the k^{th} observed edge with endpoint degrees of e_{k_m} and e_{k_f} for node types male and female, respectfully. Let $V(A)$ represent the vectorized form of a matrix A, where the columns are stacked one on top of the other. Finally, let X_k denote a matrix of the k^{th} observation such that $X_{k_{\{i,j\}}} = \begin{cases} 1 \text{ if } e_{k_m} = i \text{ and } e_{k_f} = j \\ 0 \text{ otherwise.} \end{cases}$

In many study designs $\pi_{(i,j)}$ is not a constant for all i and j, thus $\lim_{n\to\infty} 1/n \sum_k V(X_k) \nrightarrow v(\mu)$. A Horvitz-Thompson estimator can be used to construct an unbiased estimator, $\hat{\mu}_{HT}$, by reweighing the observations. In the study design presented in section 3, $\pi_{(i,j)} = \alpha*(1/i+1/j)$, where α is the normalizing constant. Let $V(\hat{\mu}_{HT}) = A_{HT} * 1/n \sum_k V(X_k)$ where $A_{HT} = diag(\pi_{(1,1)}, \pi_{(2,1)}, \cdots, \pi_{(F_0,M_0)})$ and F_0 (M_0) represent the maximum female (male) degree.
In the setting of sexual networks, adjacent cells in the degree mixing matrix are positively related to one another, we expect the true degree mixing matrix to be relatively smooth. But in practice only a small percentage of nodes are sampled, and this can cause our estimator to have large jumps between adjacent entries as well as a large number of zero entries. Additional issues are that the estimated degree mixing matrix may not reflect our known degree distribution, nor be graphical as defined in section 2. To address these issues, we will apply a local linear smoother, optimize over all matrices that are graphical and fit our known degree distribution.

3.3.1 Local Linear Smoothing

To smooth our estimator, we use a local linear estimator (Simonoff, 1996). For entry (i, j), this estimator will be denoted $\hat{\mu}_{LL_{(i,j)}}$, which equals $\hat{B}_0$, where $\hat{B}$ is the minimizer of

$$\sum_{k,l} [\hat{\mu}_{HT_{(i,j)}} - B_0 - B_1 * (\frac{i}{F_0} - \frac{k}{F_0}) - B_2 * (\frac{j}{M_0} - \frac{l}{M_0})]^2 * W_{h_{F_0},h_{M_0}}(i, j, k, l, F_0, M_0).$$

Let $W^{(i,j)} = diag[W_{h_i,h_j}(\frac{i-1}{F_0}, \frac{j-1}{M_0}), W_{h_i,h_j}(\frac{i-2}{F_0}, \frac{j-1}{M_0}), \cdots, W_{h_i,h_j}(\frac{i-F_0}{F_0}, \frac{j-M_0}{M_0})]$, and

$$X^{(i,j)^T} = \begin{pmatrix} 1 & 1 & & 1 & 1 & & 1 \\ \frac{i-1}{F_0} & \frac{i-2}{F_0} & \cdots & \frac{i-F_0}{F_0} & \frac{i-1}{F_0} & \cdots & \frac{i-F_0}{F_0} \\ \frac{j-1}{M_0} & \frac{j-1}{M_0} & \cdots & \frac{j-1}{F_0} & \frac{j-2}{M_0} & \cdots & \frac{j-M_0}{M_0} \end{pmatrix}.$$

Define $\hat{B}_x = \left(\left\{ (X^{(i,j)})^T * W^{(i,j)} * X^{(i,j)} \right\}^{-1} (X^{(i,j)})^T * W^{(i,j)} \right)$. Thus, $\hat{B} = \hat{B}_x * V(\hat{\mu}_{HT})$ and $\hat{\mu}_{LL_{(i,j)}} = W_x^{(i,j)} * V(\hat{\mu}_{HT})$, where $W_x^{(i,j)}$ is the first row of $\hat{B}_x$. Let $V(\hat{\mu}_{LL}) = A_{LL} * V(\hat{\mu}_{HT})$ where $A_{LL^T} = \begin{pmatrix} W^{(1,1)}, & W^{(2,1)}, & \cdots & , W^{(F_0,M_0)} \end{pmatrix}$.
A common choice for a two-dimensional kernel function, $W_{h_i,h_j}(\frac{i-k}{F_0}, \frac{j-l}{M_0})$, is a bivariate normal. We purpose using the following for the kernel function with smoothing parameters, h_i and h_j:

$W_{h_i,h_j}(\frac{i-k}{F_0}, \frac{j-l}{M_0}) = N_2\left(\phi = (i,j), \Sigma = \begin{pmatrix} h_i & 0 \\ 0 & h_j \end{pmatrix}\right)$.

In many sampling designs, including the one proposed in section 3, it is common for the bottom right section of observed degree mixing matrix to have significantly fewer observations than the rest of the matrix, making additional smoothing necessary for those entries. To achieve such smoothing, we can either transform the data, or allow the smoothing parameters to depend on the total number of observations for each row and column. In practice, the latter appeared to work well, and we choose smoothing parameters to be $h_i = 1/\sqrt{n_{r_i}}$ and $h_j = 1/\sqrt{n_{c_j}}$, where n_{r_i} and n_{c_j} are the total number of observations in row i and column j, respectively.

3.3.2 Linear Programming

The next step is ensuring that the matrix $\hat{\mu}_{LL}$ fits as closely as possible the degree percentages observed in the network. To accomplish this goal, we will use a linear programming framework to compute the weighting matrix B, where the final estimate will be $V(\hat{\mu}_{LP}) := A_{LP} * V(\hat{\mu}_{LL}) := (B + I) * V(\hat{\mu}_{LL})$. Minimize $\sum_l \sum_m |B_{lm}| * V(\hat{\mu}_{LL})_m + \sum_l \sum_m |B_{lm}|$ is proposed for the linear programming objective function. The first sum limits the magnitude of any reweighing and the second ensures that no weight is extremely large. To ensure that the estimated degree mixing matrix marginals equal the known degree distribution we require the following linear constraints to hold: $\hat{\mu}_{LP_{(1,k)}} + \cdots + \hat{\mu}_{LP_{(H_0,k)}} = D_k^g / \sum_j j * D_j^g$ where g and h are distinct node types and $k \in \{1, \cdots, d_g\}$. The equality: $V(\hat{\mu}_{LP})_l = B_{l1} * V(\hat{\mu}_{LL})_1 + B_{l2} * V(\hat{\mu}_{LL})_2 + \cdots + B_{l,F_0*M_0} * V(\hat{\mu}_{LL})_{F_0*M_0} + V(\hat{\mu}_{LL})_l$ allows any constraint to be written in the canonical form, $C * B = RHS$, for linear programming problems. In addition to the marginal constraints we also include non-negativity constraints on all entries in $\hat{\mu}_{LP}$, $B_{l1} * V(\hat{\mu}_{LL})_1 + B_{l2} * V(\hat{\mu}_{LL})_2 + \cdots + B_{l,F_0*M_0} * V(\hat{\mu}_{LL})_{F_0*M_0} \geq -V(\hat{\mu}_{LL})_l$.

Theorem 3.1 states that no additional constraints are needed for the estimated matrix to be graphical, if the size of the population is flexible. For the estimation of a non-bipartite network we would need to include the additional constraints outlined in Theorem 3.1.

Proposition 3.2 *$\hat{\mu}_{LP}$ is a consistent estimator for μ.*

The proof of proposition 3.2 is located in Appendix G.

3.4 Comparison

Our method was tested by using data on a sexual network from Likoma Island, which is the most complete sexual network ever collected for Sub-Saharan Africa. The network has 3661 nodes and 2801 edges (Helleringer and Kohler, 2007); its geographical location allows us to evaluate our method on a population much affected by HIV. In our analysis, individuals with no sexual partners were excluded; any sexual relation claimed by either of the partners was included. Below we consider both estimation of the degree mixing matrix and network construction.

3.4.1 Estimation

Validation of the estimation procedure described in section 4 is based on comparison of the estimated and true degree mixing matrices. As no standard method exists for such comparison, we consider two possible approaches. The first computes the assortativity statistic, $L(\mu) = \sum_i \sum_j i * j * \mu_{(i,j)}$, described in Li et al. (2004). The second sums the absolute differences of each entry of the matrices, $\sum_i \sum_j |\mu_{ij} - \hat{\mu}_{ij}|$. Figure 3.2 shows how each estimation stage, Horvitz-Thompson Adjustment (HTA), Local Linear Smoothing (LL), and Linear Programming (LP), performs in regard to these two metrics.

Figure 3.2 demonstrates that even with a small percentage of sampled nodes ($\leq 5\%$) followed by a random edge trace, the estimate of the degree mixing matrix can be accurately measured in a network the size of Likoma Island. Using an initial sample of 5% of the nodes, the average percent difference between the sampled and true assortativity statistics is only 3.3% after the smoothing procedure and linear programming described in section 4 are applied. These procedures greatly increase the precision of the estimate compared to use of the Horvitz-Thompson adjustment alone.

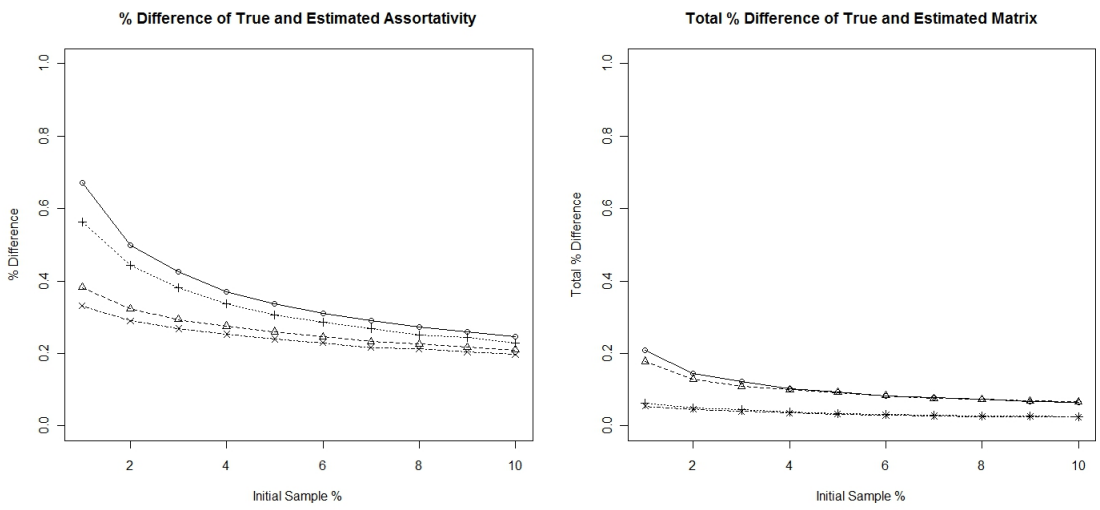

Figure 3.2: The first chart shows the percentage difference between each estimate (HTA - circle, HTA+LL - triangle, HTA+LP - plus, and HTA+LL+LP - x) and the true value for the assortativity statistic $L(\mu)$ for samples of 1-10% of the nodes in Likoma Island. The second charts shows the total percent difference for our second statistic, $\sum_i \sum_j |\mu_{ij} - \hat{\mu}_{ij}|$. 1000 simulations were run for each sample percentage.

3.4.2 Network Construction

To evaluate how network construction using a estimate for degree mixing limits the space of network topologies, we compare network properties from networks generated from the degree mixing matrix distribution as outlined above to those from networks with a uniform degree mixing matrix distribution; in both cases the marginals are treated as known. A uniform distribution on the degree mixing matrix reflects upon an assumption that individuals form partnerships based on the partner's degree, but that no information is known about the frequency of this mixing.

Fifty simulations were conducted to understand the informative degree mixing distribution, each of which began by sampling 5% of the nodes as described in section 3.4. The estimation procedure is performed on each sample, S_i, to obtain an estimated mean, $\hat{\mu}_i$, and variance, $\hat{\Sigma}_i$, for the multivariate normal distribution on degree mixing matrices. The MCMC algorithm was used to generate a chain of 6,000,000 networks for each simulation. The first 1,000,000 were discarded for MCMC burn-in. Of the remaining 5,000,000 networks, every thousandth network was used to calculate network properties. Fig-

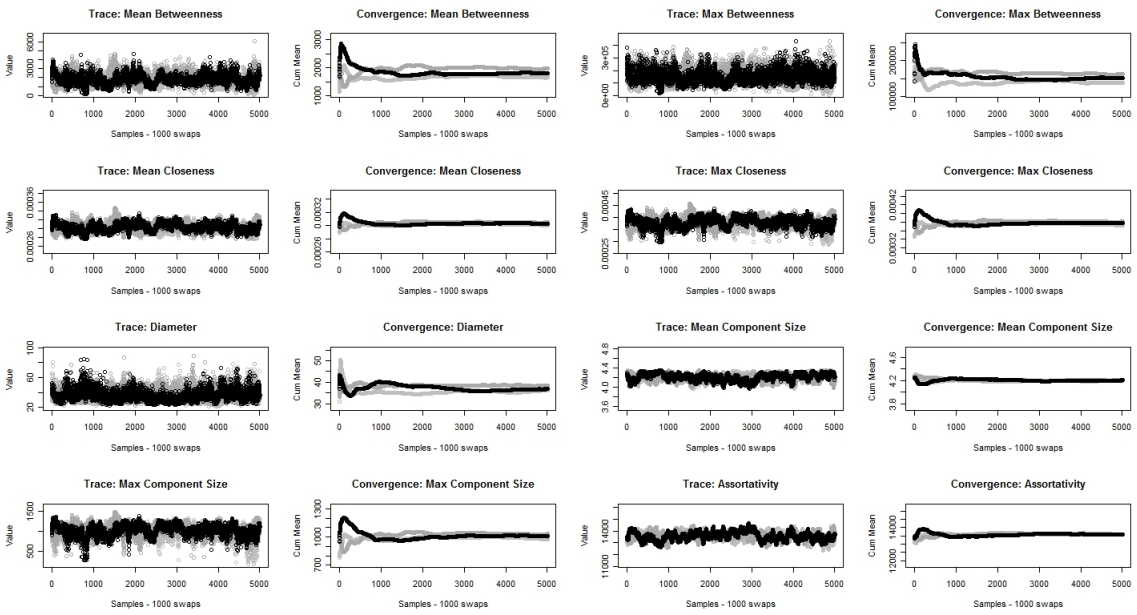

Figure 3.3: Convergence Plots: Trace and Cumulative means for eight network properties starting at multiple locations

ure 3.3 shows convergence plots of our MCMC algorithm by starting the chain at multiple locations.

Figure 3.4 demonstrates a typical network construction simulation from a single 5% random sample of nodes followed by a random trace of an edge for each node sampled in the Likoma Island dataset. The figure contains density plots using estimated degree mixing matrix distribution versus uniform degree mixing matrix distribution with known marginal. The lighter curves represent the density using the estimated degree mixing matrix. The darker curves are using a uniform (non-informative) distribution on the degree mixing matrix (using only ego-centric degree distribution). 100,000,000 graph were generated where every thousandth network was used to calculate network properties. The black bar is the truth from the Likoma island data. The first row of plots shows the mean and max of two different centrality measures, betweenness and closeness. The second row shows diameter, mean and max component size, and assortativity.

Table 3.1 provides a summary of 50 initial samples of 5% of nodes in the Likoma Island dataset, comparing compares various graph properties between graphs constructed from a sampled degree mixing matrix and those from a uniform degree mixing distribution.

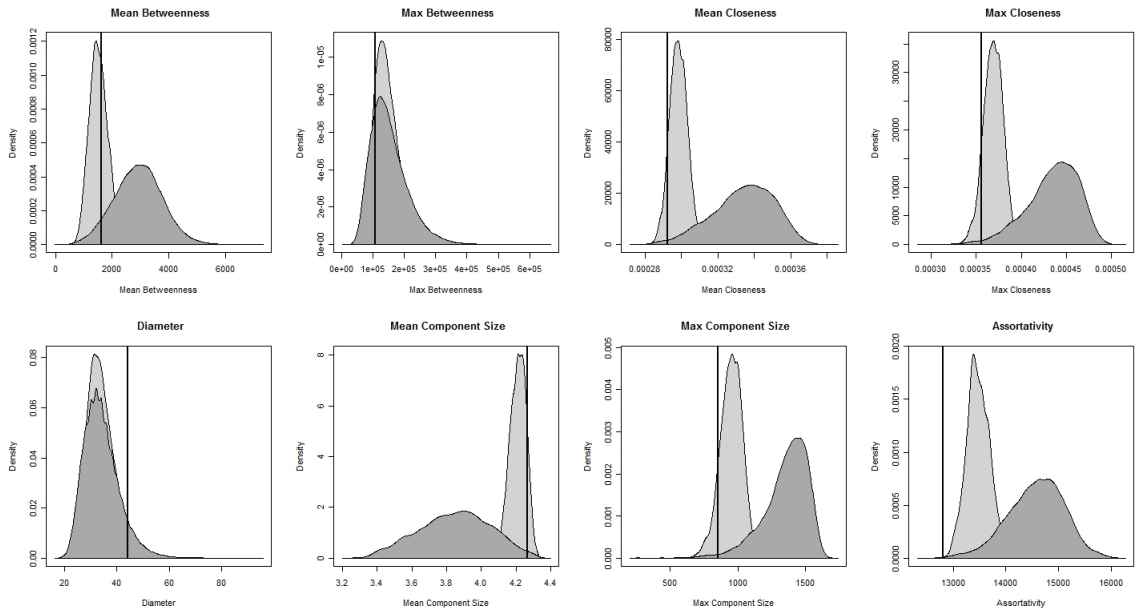

Figure 3.4: Lighter curves represent the density using the estimated degree mixing matrix. Darker curves are based on a uniform (non-informative) distribution on the degree mixing matrix over all networks satisfying the known degree sequence. The black bar is the truth from the Likoma Island data. The first row of plots show mean and max of two different centrality measures, betweenness and closeness. The second row shows diameter, mean and max component size, and assortativity.

The first column provides the percent difference between sampled networks and the truth; the second, provides the percent difference between networks with uniform distribution and the truth. The third column is the percentage of graphs constructed using the sample which are closer to the true value compared with graphs from a uniform degree mixing distribution

Table 3.1: **A Comparison of Network Properties**

Network Property	% Different Sample	% Different Uniform	% Sample Closer to Truth
Mean Betweenness	37.83	83.25	78.95
Max Betweenness	34.05	50.29	58.32
Mean Closeness	2.86	14.22	95.85
Max Closeness	5.72	22.25	95.10
Diameter	30.70	24.37	38.41
Mean Component Size	2.57	9.84	91.52
Max Component Size	18.76	58.75	93.64
Assortativity	5.95	13.75	90.61

The results from Table 3.1 show that initially sampling only 5% of nodes can provide large improvements over using ego-centric data alone. By using degree mixing information, the constructed networks are closer to the true value over 90% of the time in many network properties, as seen in Table 3.1. All the network properties listed in Table 3.1, except for assortativity, are features that were not estimated and that the network construction did not control. Table 3.1 provides evidence that degree mixing influences a range of global network properties, and therefore, without an implicit distribution for degree mixing there would be a wide range of results from network disease simulations.

3.5 Discussion

This paper presents a sampling design and estimation procedure to characterize the frequency between pairings of individuals based on their degrees, or, in our application, numbers of sexual partners within a unit of time. We showed that sampling a fairly small fraction of the population under a simple design can provide reasonably adequate information for accurate estimation of the degree mixing matrix, and that such estimates

have implications for more global properties of the network. We note, however, that networks that are actually observable may not be a random collection among all of those that can exist; selection factors may mediate the impact of knowing the degree mixing matrix on reducing the variability in other degree properties. Our estimation procedure makes use of multiple steps; the smoothing step can be achieved as described in section 4 or from regression models. The network construction procedure we propose can easily be extended to include any number of additional covariates, besides nodal degree, that can be summarized in a mixing matrix.

The estimation assumes that the degree distribution at a population level is known, but this assumption can be relaxed by using methods proposed in Appendix D. In this paper we did not discuss the issue of reporting error, which may arise even in fairly simple sampling designs. Further work is needed to understand the robustness of the estimation and construction in the presence of reporting error, and to incorporate important additional network features, such as spatial characteristics, in construction of networks.

Acknowledgements

The authors would like to thank Stéphane Helleringer and Hans-Peter Kohler for making available the Likoma Island data and their guidance in understanding the information and interpreting results. We would also like to thank Ted Cohen, Nicholas Christakis, and James O'Malley for their useful comments. This research is supported by grants from the National Institute of Health (T32AI07358, ROI AI 51164). Conflict of Interest: None declared.

4.1 Appendix A: Characterization of Valid Degree Mixing Matrices

Theorem 1.1: *Let $TDMM_{i,j}$ represent the number of edges connecting nodes of degree i to nodes of degree j. Let $TD_i = (\sum_j TDMM_{i,j} + TDMM_{i,i})/i$ (this will represent the number of nodes with degree i). An square matrix, TDMM, of dimension r is graphical by a simple undirected network if and only if the following five conditions are met.*

1. *TD_i is a non-negative integer*
2. *$TDMM_{i,j} \leq TD_i * TD_j$ if $i \neq j$*
3. *$TDMM_{i,i} \leq TD_i * (TD_i - 1)/2$*
4. *$TDMM_{i,j} \geq 0$*
5. *$TDMM_{i,j} = TDMM_{j,i}$ (symmetric)*

Before we can prove Theorem 1.1, we first need the following lemma.

Lemma: *Let $|E| \in \{0, \cdots, n * (n-1)/2\}$ where n is the number of nodes in a graph. The degree sequence d where $d_i \in \{\alpha, \alpha + 1\}$ for all $i \in \{1, \cdots, n\}$ and $\sum_{i=0}^{n} d_i = 2 * |E|$ is graphical.*

Proof of Lemma: By strong induction on $|E|$.

Base Case: $|E| = 1$. Thus, d has a size, n, ≥ 2. So d is a set of $n-2$ 0's and exactly two 1's. d is clearly graphical by creating n nodes with the last two having an edge between them.

Induction Step: Assume true for $|E| \leq N$ show for $|E| = N + 1$. Let d be a degree sequence, of size n, where $d_i \in \{\alpha, \alpha + 1\}$ for all $i \in \{1, \cdots, n\}$, $\sum_{i=0}^{n-1} d_i = 2 * (N + 1)$ and $N+1 \in \{0, \cdots, n*(n-1)/2\}$. Let $M = min\{i : d_i \geq d_j \text{ for all } j \in \{1, \cdots, n\}\}$. Let d' be equal to d except d_M is removed, thus d' is of size $n-1$. Let $\{d'_{i_1}, \cdots, d'_{i_k}\}$ be the largest d_M values in d'. Let $d'_{i_j} = d'_{i_j}$ for all $j \in \{1, \cdots, d_M\}$. This is possible because $d'_M = \lceil \frac{2*(N+1)}{n} \rceil \leq \frac{n*(n-1)}{n} = n - 1$.

In order to check if d' is graphical, we need to ensure $\frac{\sum_{i=0}^{n-1} {d'}_i}{2} \in \{0, \cdots, (n-1)*(n-2)/2\}$ and ${d'}_i \in \{\alpha, \alpha+1\}$ for all $i \in \{1, \cdots, n-1\}$. By assumption we know that $N+1 \leq \frac{n*(n-1)}{2}$. Thus, it can be shown that $N+1-\frac{2*(N+1)}{n} \leq \frac{(n-1)*(n-2)}{2}$. Since $N+1-\lceil\frac{2*(N+1)}{n}\rceil = \frac{\sum_{i=0}^{n-1} {d'}_i}{2}$ we get the desired result that $\frac{\sum_{i=0}^{n-1} {d'}_i}{2} \in \{0, \cdots, (n-1)*(n-2)/2\}$. $d'_i \in \{\alpha, \alpha+1\}$ for all $i \in \{1, \cdots, n-1\}$ is guaranteed since we are subtracting one for the degrees with the highest values and d originally had this property.

With these two conditions met, we can use the induction assumption, and thus d' is graphical. Including an isolate node at position M would still make the sequence graphical. Finally, connecting the isolate node to $\{{d'}_{i_1}, \cdots, {d'}_{i_k}\}$ would still be graphical. This new graph would have the degree sequence of d, and so d is graphical.

■

Proof of Theorem 1.1: Given an undirected graph, it is clear that the degree mixing matrix will satisfy the conditions in Theorem 1.1. Thus, we need only show that a matrix which satisfies the six criteria is graphical, which will be shown by constructing a realization of the matrix. We begin by generating an empty network with $\sum_i TD_i$ nodes, where TD_i of them will have degree i. The first condition guarantees that TD_i is a non-negative integer. To next step is adding edges to the empty graph. This will be separated into two steps. The first step is adding edges between nodes with the same final degree and the second step is adding edges between nodes with different final degrees.

Step 1: Edges between nodes with same final degree

The goal of step one is to connect $TDMM_{i,i}$ edges between nodes with final degree i for each $i \in \{1, \cdots, r\}$. We want to connect the edges such that at the end of this step each node of final degree i has one of two possible degree values, $\lfloor\frac{TDMM_{i,i}}{TD_i}\rfloor$ and $\lceil\frac{TDMM_{i,i}}{TD_i}\rceil$, for its current degree, ie the edges are added to balance the current degree as much as possible. The assignment of Deg_i ensures that maximum degree after this step, $\lceil\frac{TDMM_{i,i}}{TD_i}\rceil$, is less than the desired final degree, i. In order to prove edges can be added to maintain

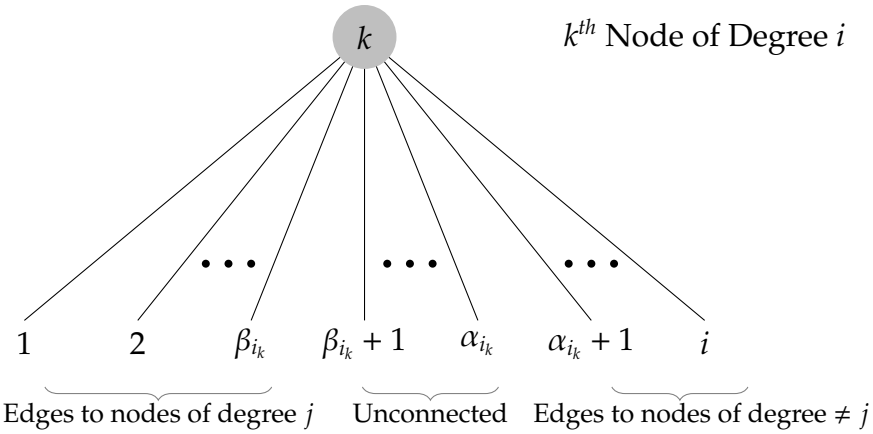

Figure 4.1: Edge connections for a node of degree i

the required degree balance we will use the lemma, where $TDMM_{i,i}$ and TD_i represent $|E|$ and n respectively. To apply the lemma, we need to insure $TDMM_{i,i} \in \{0, \cdots, \frac{(TD_i * (TD_{i-1}))}{2}\}$, which is guaranteed by condition (3) and (4).

Step 2: Connect nodes with different degrees

Once edges have been added to nodes with the same final degree, we have to add edges between nodes of degree i to nodes of degree j, for each $i, j \in \{1, \cdots, r\}$. Define the following for each i, j pair where $i \neq j$. Let $\vec{\alpha_i}$ where α_{i_k} equals i minus current degree of the k^{th} node with degree i, ie the number of edges still needed for each node. Similarly, define $\vec{\alpha_j}$ for nodes with degree j. Without loss of generality we will assume that $\vec{\alpha_i}$ and $\vec{\alpha_j}$ are in decreasing order. Define $\vec{\beta_i}$ such that $\beta_{i_k} \in \{\lfloor \frac{D_{ij}}{TD_i} \rfloor, \lceil \frac{TDMM_{i,i}}{TD_i} \rceil\}$, $\sum_k \beta_{i_k} = TDMM_{i,i}$, and $\beta_{i_1} \geq \beta_{i_2} \geq \cdots \geq \beta_{i_{TD_i}}$. β_{i_k} represents the number of edges that will be added which connect the k^{th} node with degree i with nodes of degree j. Similarly, define $\vec{\beta_j}$ for nodes with degree j. Figure 4.1 graphical describes the edge connections for a node of degree i. Connect the first degree i node to the first β_{i_1} nodes of degree j. Next connect the second degree i node to the next β_{i_2} nodes of degree j (may need to loop back to the first degree j node). This process is described in figure 4.2.

Repeat this process for all TD_i degree i nodes. This process can fail in one of three ways

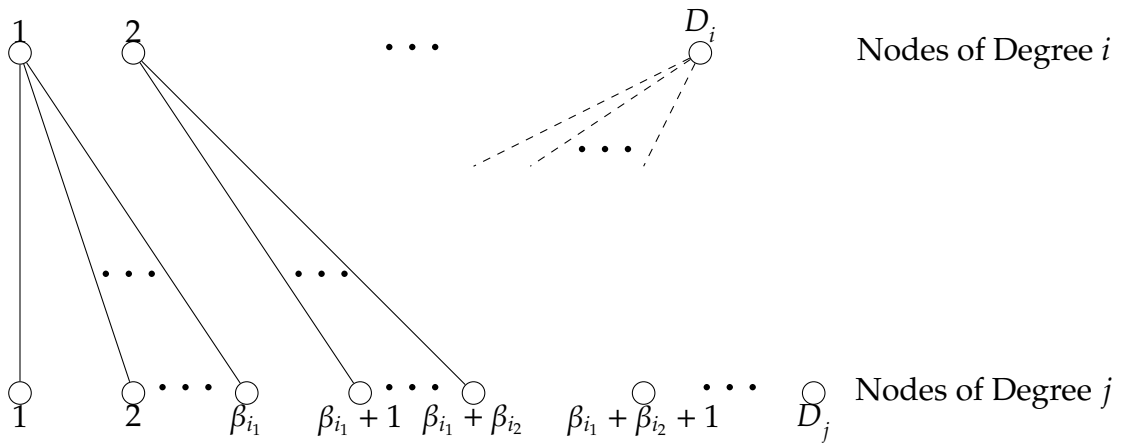

Figure 4.2: Edge connections between nodes of degree i and degree j.

to construct a graph with the degree mixing matrix of D.

Issue 1: $\beta_{i_k} > TD_j$.

The issue 1 occurs when a single node, k, of degree i must to connect β_{i_k} nodes of degree j, but β_{i_k} is greater than the number of nodes of degree j, TD_j. Thus, node k must form two edges with the same node of degree j. This cannot occur because $\beta_{i_k} \leq \lceil \frac{D_{ij}}{TD_i} \rceil \leq TD_j$ by our condition (2).

Issue 2: $\alpha_{i_k} < \beta_{i_k}$.

The second issue occurs when $\alpha_{i_k} < \beta_{i_k}$, i.e. a node of degree i has fewer unconnected edges than the number of nodes of degree j to which it is assigned. Initially when constructing the graph we generated $TD_i = (\sum_j TDMM_{i,j} + TDMM_{i,i})/i$ nodes of degree i, which means the sum degree of all the degree i nodes is $(\sum_j TDMM_{i,j} + TDMM_{i,i})$. The number of unconnected edges after step 1 is $(\sum_j TDMM_{i,j} + TDMM_{i,i}) - 2 * TDMM_{i,i} = \sum_{j \neq i} TDMM_{i,j}$. Thus, there are enough unconnected edges from nodes of degree i to connect the required number of edges, $TDMM_{i,j}$. Hence, we know $\sum_k \alpha_{i_k} - \beta_{i_k} \geq 0$. Thus, there exists partitions, p_1 and p_2, of size k of the values $\sum_k \alpha_{i_k}$ and $\sum_k \beta_{i_k}$ such that $p_{1_l} \geq p_{2_l}$ for each

$l \in \{1, \cdots, k\}$. One such pair of partitions is where each partition is decreasing and is as balanced as possible. This is exactly the partition generated under this construction proof. Throughout the construction the number of available edges for nodes with the same degree are as balanced as possible. The first step of connecting edges between nodes with the same degree initially forces this condition. In subsequent steps of connecting nodes with different degrees ensures this condition remains by assigning more edges to those nodes with more available edges. Thus, by construction $\alpha_{i_k} < \beta_{i_k}$ is not possible.

Issue 3: $\alpha_{j_k} < \beta_{j_k}$.

Due to the symmetry of i and j, the proof that $\alpha_{j_k} < \beta_{j_k}$ is not possible is identical to issue 2. ■

4.2 Appendix B: Alternative Measures of Concurrency

Alternative definitions of concurrency can be calculated using the fixed degree distribution and the values of the three measures of concurrency. Cumulative concurrency is also defined as the proportion of multiple partnerships that are concurrent in the past year Fishel et al. (2012), which can be evaluated as $\frac{C_1}{n*\sum_{i=2} D_i}$, where n is the number of individuals in the population. $\frac{C_3}{T}$ is equal to the mean point prevalent of concurrency in a population assuming that individuals have at most one concurrent relationship in a given period of time. Also, the fraction $\frac{C_3}{C_2}$ gives the average time relationships overlap, which corresponds with when concurrency has been also defined as only long-term overlapping relationships Morris et al. (2009).

4.3 Appendix C: Proof of Theorem 2.1

Theorem 2.1: The following information characterizing relationships in a given cumulative contact network $G = (V, E)$ is necessary and sufficient to construct a subspace K such that $V^{p_1}(A) = V^{p_2}(A)\ \forall\ p_1, p_2 \in K$ and $\forall\ A \in V$.

1. Listing of overlapping edges
2. Direction of non-overlapping edges
3. Labels and directions of edges completely contained in connected components of $L(G)$ using only overlapping edges.

Proof by Induction:

Let $T = \{t_1, \cdots, t_{|E|}\}$ and $R = \{r_1, \cdots, r_{|E|}\}$ be two sets of starting times for the set of relationships, E, with identical values for the information listed above in items 1-3. Let $V^T(A)$ be the set of nodes that are reachable from A using time set T. Denote $V_i^T(A) \subseteq V^T(A)$ as the subset of nodes that are reachable from A using time set T via a path of exactly i steps. Therefore, if $B \in V_i^T(A)$ means there exists a subset of nodes, $\{B_1, \cdots, B_{i-1}\}$ such that the following is a valid reachable path under time set T: $A = B_0 \to B_1, B_1 \to B_2, \cdots, B_{i-1} \to B_i = B$. We want to show that $V^T(A) = V^R(A) \forall A \in V$. In particular, we will show that $V_i^T(A) = V_i^R(A) \forall A \in V$ and $\forall i \in \mathbb{Z}^+$. Define $I_A^T(B)$ as the infection time assuming only node A was initially infected.

Base Case: i = 1

Let $B \in V_1^T(A)$, therefore there exists a path $A = B_0 \to B_1 = B$ in time set T. Since the contact network is fixed, the edge (A, B) is also present in time set R. Thus, $A = B_0 \to B_1 = B$ is a valid path in time set R, and therefore $B \in V_1^R(A)$

Base Case: i = 2

Let $B \in V_2^T(A)$, therefore there exists a path $A = B_0 \to B_1 \to B_2 = B$ in time set T. There are three possible relationships between edges (B_0, B_1) and (B_1, B_2): (B_0, B_1) and (B_1, B_2) overlap, $e(B_0, B_1) < s(B_1, B_2)$, or $s(B_0, B_1) > e(B_1, B_2)$.

Case 1: (B_0, B_1) and (B_1, B_2) overlap. Regardless of the length of overlap, $B \in V_2^R(A)$.

Case 2: $e(B_0, B_1) < s(B_1, B_2)$. Therefore $I_A^T(B_1) < s(B_1, B_2)$. So, $B_2 = B$ is able to be infected by B_1 in time set R.

Case 3: $s(B_0, B_1) > e(B_1, B_2)$. It is not possible for $B \in V_2^T(A)$ or $B \in V_2^R(A)$.

Induction Step: Assume true for $j \leq i$ where $i > 2$

Assume if $B \in V_j^T(A)$ then $B \in V_j^R(A)$ for $j \leq i$. We want to show if $B \in V_{i+1}^T(A)$ then $B \in V_{i+1}^R(A)$. Let $B \in V_{i+1}^T(A)$, thus there exists a path $A = B_0 \to B_1, B_1 \to B_2, \cdots, B_i \to B_{i+1} = B$

using time set T. We will show that the path is also valid for time set R.

Case 1: $\exists j$ such that the edges (B_{j-1}, B_j) and (B_j, B_{j+1}) do not overlap. By assumption $B_j \in V_j^R(A)$ and $B_{i+1} \in V_{i-j}^R(B_{j+1})$ since $B_j \in V_j^T(A)$ and $B_{i+1} \in V_{i-j}^T(B_{j+1})$ and $j, i-j \leq i+1$. It only has to be shown that $I_A^R(B_j) \leq s(B_j, B_{j+1})$, since this will be identical to starting the simulation at with B_j initially infected at time $s(B_j, B_{j+1})$. Since (B_{j-1}, B_j) and (B_j, B_{j+1}) do not overlap, it must be true in time set T that $e(B_{j-1}, B_j) < s(B_j, B_{j+1})$ for the path $A = B_0 \to B_1, B_1 \to B_2, \cdots, B_i \to B_{i+1} = B$ to be valid for time set T. Therefore, by item 2, $e(B_{j-1}, B_j) < s(B_j, B_{j+1})$ must be true for time set R. Hence, $I_A^R(B_j) \leq s(B_j, B_{j+1})$.

Case 2: $\nexists$ a j such that edges (B_{j-1}, B_j) and (B_j, B_{j+1}) do not overlap. We will show that $B \in V_{i+1}^R(A)$ by contradiction. Assume $B \notin V_{i+1}^R(A)$, thus, transmission does not occur in time set R between B_i and B_{i+1}, because by assumption $V_i^T(A) = V_i^R(A)$. Since all sequential edges must overlap, therefore by item 3 we know the label and direction relating edges (B_{i-2}, B_{i-1}) and (B_i, B_{i+1}). Therefore that are three possible sub-cases: (B_{i-2}, B_{i-1}) and (B_i, B_{i+1}) overlap, (B_{i-2}, B_{i-1}) and (B_i, B_{i+1}) do not overlap and $s(B_{i-2}, B_{i-1}) > e(B_i, B_{i+1})$ or (B_{i-2}, B_{i-1}) and (B_i, B_{i+1}) do not overlap and $e(B_{i-2}, B_{i-1}) > s(B_i, B_{i+1})$.

Case 2a: (B_{i-2}, B_{i-1}) and (B_i, B_{i+1}) overlap. We want to show that $\cap_{j=i-2}^{j=i}[s(B_j, B_{j+1}), e(B_j, B_{j+1})] \neq \emptyset$, because any point $p \in \cap_{j=i-2}^{j=i}[s(B_j, B_{j+1}), e(B_j, B_{j+1})]$ allows for the instantaneous potential spread of infection from B_{i-2} to B_{i+1}. Let $[p_1, p_2] = [s(B_{i-2}, B_{i-1}), e(B_{i-2}, B_{i-1})] \cap [s(B_{i-1}, B_i), e(B_{i-1}, B_i)]$. Since the intervals are continuous, $p_1 \in \{[s(B_{i-2}, B_{i-1}), s(B_{i-1}, B_i)]\}$ and $p_2 \in \{[e(B_{i-2}, B_{i-1}), e(B_{i-1}, B_i)]\}$. Since (B_i, B_{i+1}) overlaps with both (B_{i-2}, B_{i-1}) and (B_{i-1}, B_i), $\exists q_1, q_2 \in (B_i, B_{i+1})$ such that $q_1 > p_1$ and $q_2 < p_2$. Therefore, $\exists q_3$ such that $q_3 \in [p_1, p_2]$. Hence $\cap_{j=i-2}^{j=i}[s(B_j, B_{j+1}), e(B_j, B_{j+1})] \neq \emptyset$ as required.

Case 2b: (B_{i-2}, B_{i-1}) and (B_i, B_{i+1}) do not overlap and $s(B_{i-2}, B_{i-1}) > e(B_i, B_{i+1})$. Since $I_A^T(B_{i+1}) > I_A^T(B_{i-1}) > s(B_{i-2}, B_{i-1}) > e(B_i, B_{i+1})$ which is a contradiction because $I_A^T(B_{i+1}) \in [s(B_i, B_{i+1}), e(B_i, B_{i+1})]$. Therefore, case 2b is not a possibility.

Case 2c: (B_{i-2}, B_{i-1}) and (B_i, B_{i+1}) do not overlap and $e(B_{i-2}, B_{i-1}) < s(B_i, B_{i+1})$. By assumption $B_i \in V_i^R(A)$, therefore $I_A^R(B_i) \in [s(B_{i-1}, B_i), e(B_{i-2}, B_{i-1})]$. Since $e(B_{i-2}, B_{i-1}) < s(B_i, B_{i+1})$, B_i is infected before the start of relationship (B_i, B_{i+1}). Thus, B_{i+1} is able to be infected by B_i in time set R.

4.4 Appendix D: Methods for Bipartite Networks

This section describes several common scenarios that demonstrate the capabilities of our framework to handle sampled data. Once the network space and network proposal method have been selected, only the functions $P_C(C_g)$ and $f(C_g, C_h)$ need to be specified. In all scenarios, the network space consists of all bipartite networks with a fixed degree distributions for each gender. Edge toggling is used to propose networks (described in [Paper 1]), but restricted to only edges with opposite types of nodes as endpoints, a requirement of bipartite networks. As the probability mass function, P_C, on congruence classes is set by the investigator, in this section we only derive $f(C_g, C_h)$. The next section provides examples of various probability mass functions associated with different sampling strategies. Denote $g, h \in \mathscr{G}$ as the current and proposal network, respectively. Let C_g and C_h denote the congruence classes for g and h. Let the edge, (i, j), between node i and node j be the connection that is toggled to move from g to h and back. Without loss of generality, let $(i, j) \in h$ but $(i, j) \notin g$.

4.4.1 Topological Features

Density

For density, a congruence class is set of networks with the same number of edges, since all graphs in $\mathscr{G}$ have the same number of nodes. Let $|E_g|$ denote the number of edges in graph g. Networks g_1 and g_2 are in the same congruence class if and only if $|E_{g_1}| = |E_{g_2}|$. Since $(i, j) \in h$ but $(i, j) \notin g$, $|E_h| = |E_g| + 1$. To calculate $f(C_h, C_g)$ we need to know the average number of elements in C_g that are valid proposals from any element $h \in C_h$. Since removing any edge in h will produce a graph in C_g there are exactly $|E_h|$ valid proposals in C_g from graph h, and this is true regardless of the choice of $h \in C_h$. Thus,

$$f(C_h, C_g) = |E_h| \tag{4.1}$$

To calculate $f(C_g, C_h)$, we need to know the average number of elements in C_h that are valid proposals from any element $g \in C_g$. Adding any edge in g, which does not exist, will produce a graph in C_h, hence there are exactly $(\sum_i D_i^f(g) * \sum_j D_j^f(g)) - |E_g|$ valid proposals

in C_h from graph g. Again, is it true for any $g \in C_g$. Thus,

$$f(C_h, C_g) = (\sum_i D_i^f(g) * \sum_j D_j^f(g)) - |E_g| \tag{4.2}$$

Degree Distribution

For degree distribution, congruence classes are sets of networks with identical numbers of nodes and degree distribution. Thus, networks g_1 and g_2 are in the same congruence class if and only if $D_k^f(g_1) = D_k^f(g_2)$ and $D_k^m(g_1) = D_k^m(g_2)\ \forall\ k$. As g only differs from h through a toggling of the edge (i, j) , $D_k(g) = D_k(h)$ for all k except possibly $k = d_i^m(g), d_j^f(g), d_i^m(h)$ and $d_j^f(h)$. Since the only difference between the graph g and h is edge $(i, j) \in h$ but $(i, j) \notin g$, $d_i^m(h) = d_i^m(g) + 1$ and $d_j^f(h) = d_j^f(g) + 1$.

The number of edge toggles from a graph $h \in C_h$ to any graph in C_g is equal to the number of edges in h that have endpoint degrees of $d_i^m(h)$ and $d_j^f(h), DMM_{d_i^m(h), d_j^f(h)}$. Thus, $f(C_h, C_g)$ is equal to the average of $DMM_{d_i^m(h), d_j^f(h)}$ over all graphs $h \in C_h$. Let $E(DMM|C_h)$ denote the expected degree mixing matrix over graph that are in C_h. Since $h' \in C_h$ if and only if $D^m(h') = D^m(h)$ and $D^f(h') = D^f(h), E(DMM|C_h) = E(DMM|D^m(h), D^f(h))$. Thus,

$$f(C_h, C_g) = E(DMM_{d_i^m(h), d_j^f(h)}|D^m(h), D^f(h)) \tag{4.3}$$

Following arguments from (Newman, 2002) [Paper 1], based on the probability that a node's neighbor will have degree k is proportional to $k * D_k$ and not D_k,

$$E(DMM_{x,y}|D^m(h), D^f(h)) \approx \frac{D_x^m * x * D_y^f * y}{\sum_k D_k^m * k} \tag{4.4}$$

The number of edge toggles from a graph $g \in C_g$ to any graph in C_h is equal to the number of possible non-loop edges with the male endpoint degree as $d_i^m(g)$ and female endpoint degree as $d_j^f(g)$ minus the number of edges that will generate a multi-edge. The expected number of edge toggles that generate a multi-edge is $E(DMM_{d_i^m(g), d_j^f(g)}|D^m(g), D^f(g))$, denote this value as α_1.

$$f(C_h, C_g) = D_{d_i^m(g)}(g) * D_{d_j^f(g)}(g) - \alpha_1. \tag{4.5}$$

Degree Mixing and Degree Distribution

We consider a partition of $\mathscr{G}$ such that networks g_1 and g_2 are in the same congruence class if and only if $D_x^m(g_1) = D_x^m(g_2)$ and $D_x^f(g_1) = D_x^f(g_2) \forall x$ and $DMM_{x,y}(g_1) = DMM_{x,y}(g_2) \forall x, y$. Since the degree mixing matrix uniquely determines the degree distributions for males and females, an identical partition is defined when networks g_1 and g_2 are in the same congruence class if and only if $DMM(g_1) = DMM(g_2)$. Thus, the probability mass function can be defined using the degree mixing matrix. Similar to degree distribution, the number of edge toggles from a graph $h \in C_h$ to any graph in C_g is equal to $f(C_h, C_g) = DMM_{d_i^m(h), d_j^f(h)}$, since, in this setting, $E(DMM|C_h) = DMM(h)$ because all graphs in C_h have the same DMM. Similar logic holds for $f(C_g, C_h)$, thus by substituting the true degree mixing matrix for the expected degree mixing matrix in equations (5) and (7), we get the following expressions for $f(C_g, C_h)$ and $f(C_h, C_g)$.

$$f(C_h, C_g) = DMM_{d_i^m(h), d_j^f(h)}(h). \tag{4.6}$$

$$f(C_g, C_h) = D_{d_i^m(g)}(g) * D_{d_j^f(g)}(g) - \alpha_2. \tag{4.7}$$

where $\alpha_2 = DMM_{d_i(g), d_i(g)}(g)$.

As with the degree distribution, not all degree mixing matrices have a valid realization. Appendix E provides a method to characterize valid degree mixing matrices. Using the construction procedure in the Appendix E to set the initial network with the estimated degree distribution and degree mixing will tend to decrease time to convergence in the MCMC procedure.

4.4.2 Nodal Covariates

The methods developed for topological network features can be extended to include mixing patterns based on nodal covariates. Let p be the number of distinct nodal covariate

patterns of interest in the population. The covariate patterns can represent single or multiple nodal characteristics. We describe a common scenario in which we observe not only mixing patterns between covariate patterns but also the degree distributions of males and females, $\{D^{m,1}, \cdots, D^{m,p}\}$ and $\{D^{m,1}, \cdots, D^{m,p}\}$, for each covariate pattern. The following approach can be simplified for settings wherein individual covariate pattern degree distributions are not observed. In order to incorporate covariate information, knowledge of the number of individuals with covariate pattern k for each gender, M_k^m and M_k^f, is required for each k.

Nodal Covariate Mixing and Degree Distribution

For nodal covariate mixing and degree distribution, the congruence classes contain networks with identical numbers of nodes, degree distributions, and nodal covariate mixing matrices. Thus, networks g_1 and g_2 are in the same congruence class if and only if $D_x^{m,k}(g_1) = D_x^{m,k}(g_2)$ and $D_x^{f,k}(g_1) = D_x^{f,k}(g_2) \forall x, k$ and $MM_{k,l}(g_1) = MM_{k,l}(g_2) \forall k, l$.
For each covariate degree distribution, one entry in the mixing matrix is fixed. Therefore, given degree distribution estimates for each of the covariate patterns, the probability mass function can only be specified for the degree distributions and the entries above the diagonal in the mixing matrix. As above, expected degree mixing matrices, $E(DMM^{k,l})$, are constructed for each entry in the upper triangle of the covariate mixing matrix, thus $k \neq l$. The matrix entry $DMM_{x,y}^{k,l}(g)$ represents the percentage of edges where male node has covariate pattern k and degree x, while the female node has covariate pattern l and degree y. Using the setup from the previous section, we let the edge set of g and h be identical except that $(i, j) \notin g$ and $(i, j) \in h$. Regarding covariate information, let nodes i and j have covariate patterns m_i^m and m_j^f, respectively. The number of edge toggles from a graph $h \in C_h$ to any graph in C_g is equal to the number of edges in h where one endpoint has degree $d_i^m(h)$ and type m_i^m and the other endpoint has degree $d_j^f(h)$ and type m_j^f. The proportion of edges where both endpoints are specified as type m_i^m and type m_j^f compared to edges where one endpoint is required to be of type m_i^m is $\frac{MM_{i,j}}{\sum_z MM_{i,z}}$. Using similar arguments as above we can calculate the expected degree mixing matrix where only edges between types m_i^m and m_j^f are considered.

$$E(DMM^{k,l}_{x,y}|D^{m,k,l}, D^{f,k,l}) \approx \frac{D^{m,k,l}_x * x * D^{f,k,l}_y * y}{\sum_z D^{m,k,l}_z * z} \tag{4.8}$$

where,

$$D^{(m,k,l)} = D^{m,k} * \frac{MM_{k,l}}{\sum_z MM_{k,z}} \tag{4.9}$$

and

$$D^{(f,k,l)} = D^{f,l} * \frac{MM_{k,l}}{\sum_z MM_{z,l}} \tag{4.10}$$

Thus,

$$f(C_h, C_g) = E(DMM^{m_i,m_j}_{d_i(h),d_j(h)}(h)|m, D^{m_i,m_j}(h), f, D^{m_i,m_j}(h)) * |E^{k,l}_h| \tag{4.11}$$

and,

$$f(C_g, C_h) = M^m_{m_i} * M^f_{m_j} - \alpha_3 \tag{4.12}$$

where $\alpha_3 = E(DMM^{m_i,m_j}_{d_i(g),d_j(g)}(g)|D^{m_i,m_j}(g), D^{m_i,m_j}(g))$.

Nodal Covariate Mixing, Degree Mixing, and Degree Distribution

In a similar fashion as above the proposed method can be extended to include degree mixing. Once again, we substitute the true degree mixing matrices for the expected degree mixing matrices.

4.5 Appendix E: Proof of Theorem 3.1

Theorem 3.1: *An matrix, DMM, is graphical by a bipartite undirected network if and only if the following four conditions are met given degree distributions D^m and D^f:*

1. $D^f_i := (\sum_j DMM_{(i,j)})/i \in \mathbb{Z}^+ \forall i$
2. $D^m_j := (\sum_i DMM_{(i,j)})/j \in \mathbb{Z}^+ \forall j$

3. $DMM_{(i,j)} \leq D_i^f * D_j^m$

4. $DMM_{(i,j)} \geq 0$

Proof of Theorem 3.1: Given a undirected bipartite graph it is clear that the degree mixing matrix will satisfy the conditions in Theorem 3.1. Thus, we need only show that a matrix which satisfies the four criteria is graphical via a bipartite graph. As with Theorem 1.1 of chapter 1, this will be shown by constructing a realization of the matrix. We begin by generating an empty network with $\sum_i D_i^f$ and $\sum_i D_i^m$ nodes of type 1 and type 2, respectively, where D_i^f and D_i^m of type 1 and type 2 will have degree i. Conditions (2) and (3) guarantee that D_i^f and D_i^m are non-negative integers.

The next step is to add edges between type 1 nodes of degree i to type 2 nodes of degree j, for each $i \in \{1, \cdots, r\}$ and $j \in \{1, \cdots, s\}$. We will use a similar approach to connect type 1 nodes with degree i to type 2 nodes of degree j as presented in the proof for Theorem 1.1. The approach starts by defining the components of $\vec{\alpha_i^{(1)}}$, $\alpha_{i_k}^{(1)}$, as the available edges left to be connected for the k^{th} type 1 node with degree i. The components of $\vec{\beta_i^{(1)}}$ are defined so that $\beta_{i_k}^{(1)} \in \{\lfloor \frac{DMM_{(i,j)}}{D_i^f} \rfloor, \lceil \frac{DDM_{(i,j)}}{D_i^f} \rceil\}$, $\sum_k \beta_{i_k}^{(1)} = DMM_{(i,j)}$, and $\beta_{i_1}^{(1)} \geq \beta_{i_2}^{(1)} \geq \cdots \geq \beta_{i_{D_i^f}}^{(1)}$. Let $\vec{\alpha_j^{(2)}}$ and $\vec{\beta_j^{(2)}}$ be defined similarly for type 2 nodes with degree j. Without loss of generality assume that $\vec{\alpha_i^{(1)}}$ and $\vec{\alpha_j^{(2)}}$ are in decreasing order. Similar to the second step in the proof of theorem one, this construction will connect the first type 1 node of degree i to the first $\beta_{i_1}^{(1)}$ type 2 nodes in $\vec{\alpha_j^{(2)}}$. Next, we connect the second type 1 node of degree i to the next $\beta_{i_2}^{(1)}$ nodes in $\vec{\alpha_j^{(2)}}$, and repeat this process for all D_i^f degree i nodes.

Again, similar to proof of Theorem 1.1 of chapter 1, there are only three issues that could arise.

Issue 1: $\beta_{i_k}^{(1)} > D_j^m$.

The issue 1 occurs when a single node of type 1, k, of degree i must connect to $\beta_{i_k}^{(1)}$ type 2 nodes of degree j, but $\beta_{i_k}^{(1)}$ is greater than the number of type 2 nodes of degree j, D_j^m. Thus, node k must form two edges with the same node of degree j. This cannot occur

because $\beta^{(1)}_{i_k} \leq \lceil \frac{DDM_{(i,j)}}{D^f_i} \rceil \leq D^m_j$ by our condition (1).

Issue 2: $\alpha^{(1)}_{i_k} < \beta^{(1)}_{i_k}$.

Initially there is a total of $\sum_j DMM_{(i,j)}$ available edges of type 1 nodes of degree i to connect to type 2 nodes. At each step of connecting type 1 nodes of degree i to type 2 nodes of l, $DMM_{(i,l)}$ available edges are removed. Thus, at the step of connecting type 1 nodes of degree i to connect to type 2 nodes of degree j, there exists at least $D_{(i,j)}$) available edges. So, $\sum_k \alpha^{(1)}_{i_k} \geq DMM_{(i,j)} = \sum_k \beta^{(1)}_{i_k}$. Thus, if you create size D^f_i partitions, p_1 and p_2, of $\sum_k \alpha^{(1)}_{i_k}$ and $\sum_k \beta^{(1)}_{i_k}$, where the values in the partitions are decreasing and as balanced as possible, $p_1 \geq p_2$ for each pairwise element. These particular partitions are exactly what the algorithm is generating with $\alpha^{(1)}_{i_k}$ and $\beta^{(1)}_{i_k}$. So, it can be concluded that $\alpha^{(1)}_{i_k} \geq \beta^{(1)}_{i_k}$.

Issue 3: $\alpha^{(2)}_{j_k} < \beta^{(2)}_{j_k}$.

Due to the symmetry of type 1 and type 2 nodes, the proof that $\alpha^{(2)}_{j_k} < \beta^{(2)}_{j_k}$ is not possible is identical to issue 2. ■

4.6 Appendix F: Proof of Proposition 3.1

Proposition 3.1: *Given* $C_g \neq C_h$, $f(C_g, C_h) \approx DMM_{(j,k)}(g) * DMM_{(l,i)}(g) * (1 - P_1 - P_2 + P_1 * P_2)$, *where* $P_1 = \frac{(i-1)*(j-1)*DMM_{(j,i)}(g)}{(i*D^f_i-1)*(j*D^m_j-1)}$ *and* $P_2 = \frac{(k-1)*(l-1)*DMM_{(k,l)}(g)}{(k*D^f_k-1)*(l*D^m_l-1)}$.

Proof of Proposition 3.1: We calculate $f(C_g, C_h)$ under the condition that $C_g \neq C_h$. From an element in C_g there are $DMM_{(j,k)}(g) * DMM_{(l,i)}(g)$ ways to select two edges, e_1 and e_2, with endpoint degrees of j and k for e_1 and l and i for e_2. When the endpoints of e_1 and e_2 are switched it produces a graph with the same DMM as an element in C_h, but the graph may have a multi-edge, and hence not a valid proposal graph. Thus,

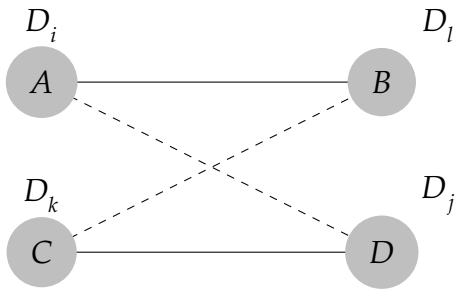

Figure 4.3: Edge switching by replacing edges (A,B) and (C,D), solid lines, with edges (A,D) and (C,B), dashed lines.

$$\begin{aligned} f(C_g, C_h) &= DMM_{(j,k)}(g) * DMM_{(l,i)}(g) - E[\text{\# of invalid moves}] && (4.13)\\ &= DMM_{(j,k)}(g) * DMM_{(l,i)}(g) - (\text{\# of switches}) * P(\text{multi-edge}) && (4.14)\\ &= DMM_{(j,k)}(g) * DMM_{(l,i)}(g) * (1 - P(\text{multi-edge})) && (4.15) \end{aligned}$$

where P(multi-edge) is the probability of an creating a multi-edge. To calculate P(multi-edge) we consider a graph $G \in C_g$ with nodes A, B, C, and D of degree i, l, k, and j, respectively. Assume there exist edges (A, B) and (C, D). Switching those edges yields a graph, $H \in C_h$, but it may contain a multi-edge. A multi-edge occurs when there already exists an edge (A, D) or (C, B) before the edge switch which is shown in the figure below.

Thus,

$$\begin{aligned} P(\text{multi-edge}) &= P((A,D) \text{ or } (C,B) | (A,B),(C,D)) && (4.16)\\ &\approx P_1 + P_2 - P_1 * P_2 && (4.17) \end{aligned}$$

where P_1 is $P((A, D)|(A, B), (C, D))$ and P_2 is $P((C, B)|(A, B), (C, D))$. In order to calculate $P((A, D)|(A, B), (C, D))$ we first need to calculate the number of edges from A to a degree j node, T_1 edges, and the number of edges from D to a degree i node, T_2 edges. Let

$$S_1 = \text{\# of } T_1 \sim Bin(i - 1, \frac{DMM_{(j,i)}}{i * D_i^f - 1}). \quad (4.18)$$

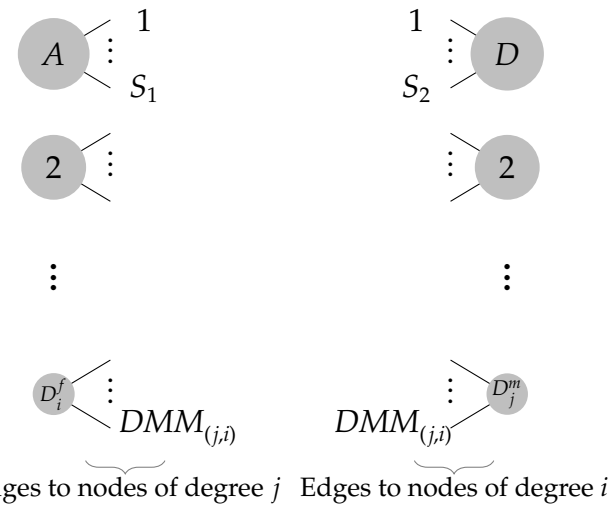

Figure 4.4: Attaching edges of Node A to Node D

There are $i * D_i^f - 1$ remaining edges from female nodes of degree i and $DMM_{(j,i)}$ of those edges must connect to a male node of degree j. Thus, each of the $i - 1$ remaining edges from node A have a $DMM_{(j,i)}/(i * D_i^f - 1)$ probability of being a T_1 edge. Similarly, let

$$S_2 = \#\text{ of } T_2 \sim Bin(j - 1, \frac{DMM_{(j,i)}}{j * D_j^m - 1}). \tag{4.19}$$

Thus, $P((A, D)|(A, B), (C, D))$ is the probability that any of the T_1 edges connects with one of the T_2 edges given there are $DMM_{(j,i)}$ options. Figure below illustrates possible edge connections from node A to node D.

Since S_1 and S_2 are small compared to D_j^m and D_i^f, each of the T_1 edges have approximately $S_2/DMM_{(j,i)}$ probability of connecting with a T_2 edge. Hence,

$$\begin{aligned} P_1 &\approx E[S_1 * S_2/DMM_{(j,i)}] = E[S_1] * E[S_2] * 1/DMM_{(j,i)} && (4.20) \\ &= \frac{(i-1) * (j-1) * DMM_{(j,i)}}{(i * D_i^f - 1) * (j * D_j^m - 1)} && (4.21) \end{aligned}$$

Similarly,

$$P((C, B)|(A, B), (C, D)) \approx \frac{(k-1) * (l-1) * DMM_{k,l}}{(k * D_k^f - 1) * (l * D_l^m - 1)}. \tag{4.22}$$

Substituting equations (3.23) and (3.24) in equation (3.19) and using that quantity in equation (3.17), we get the following:
$f(C_g, C_h) \approx DMM_{(j,k)}(g) * DMM_{(l,i)}(g) * (1 - P_1 - P_2 + P_1 * P_2)$, where $P_1 = \frac{(i-1)*(j-1)*DMM_{(j,i)}}{(i*D_i^f-1)*(j*D_j^m-1)}$ and $P_2 = \frac{(k-1)*(l-1)*DMM_{(k,l)}}{(k*D_k^f-1)*(l*D_l^m-1)}$. ■

4.7 Appendix G: Proof of Proposition 3.2

Proposition 3.2: *$\hat{\mu}_{LP}$ is a consistent estimator for μ.*

Proof of Proposition 3.2: The Horvitz-Thompson estimator (Overton and Stehman, 1995) and the local linear estimator (Simonoff, 1996) both maintain the property of consistency from the original sample. Thus, it only has to be shown that the estimator after linear programming is still consistent.
Consistency requires the following: $\lim_{n\to\infty} P(|V(\hat{\mu}_{LP}^{(n)})_i - V(\mu)_i| \geq \epsilon) \to 0$ or equivalently $\lim_{n\to\infty} P(|V(\hat{\mu}_{LL}^{(n)})_i + \sum_i B_{(j,i)}^{(n)} * V(\hat{\mu}_{LL}^{(n)})_i - V(\mu)_i| \geq \epsilon) \to 0$.
We know that $\lim_{n\to\infty} V(\hat{\mu}_{LL}^{(n)})_i \to \mu_i$, since our kernel smoothing produces consistent estimators. If we show $\lim_{n\to\infty} B_{(j,i)}^{(n)} \to 0$ for all j and i, then by Slutsky's Theorem $\lim_{n\to\infty} V(\hat{\mu}_{LL}^{(n)})_i + \sum_i B_{(j,i)}^{(n)} * V(\hat{\mu}_{LL}^{(n)})_i \to V(\mu)_i$.
To show $\lim_{n\to\infty} B_{(j,i)}^{(n)} \to 0$ for all j and i we will introduce a new set of weights, $W^{(n)}$.
Let *max* be the minimum index where the maximum value is obtained in the vector $V(\mu)$, i.e. $max = min(i : V(\mu)_i \geq V(\mu)_j)$ and let

$$W_{(j,i)}^{(n)} = \begin{cases} \dfrac{V(\mu)_j - V(\hat{\mu}_{LL}^{(n)})_j}{V(\hat{\mu}_{LL}^{(n)})_{max}} \; if \; i = max \\ 0 \text{ otherwise.} \end{cases} \tag{4.23}$$

Now it will be shown that $W^{(n)}$ is a valid solution to our linear programming problem.

$$(W^{(n)} + I) * V(\hat{\mu}_{LL}^{(n)}) = \begin{pmatrix} 1 & 0 & \cdots & 0 & W_{(1,max)}^{(n)} & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & W_{(2,max)}^{(n)} & 0 & \cdots & 0 \\ & & & & \vdots & & & \\ 0 & 0 & \cdots & 0 & W_{(d_f * d_m,max)}^{(n)} & 0 & \cdots & 1 \end{pmatrix} * \begin{pmatrix} V(\hat{\mu}_{LL}^{(n)})_1 \\ V(\hat{\mu}_{LL}^{(n)})_2 \\ \vdots \\ V(\hat{\mu}_{LL}^{(n)})_{d_f * d_m} \end{pmatrix} \quad (4.24)$$

$$= \begin{pmatrix} V(\hat{\mu}_{LL}^{(n)})_1 + W_{(1,max)}^{(n)} * V(\hat{\mu}_{LL}^{(n)})_{max} \\ V(\hat{\mu}_{LL}^{(n)})_2 + W_{(2,max)}^{(n)} * V(\hat{\mu}_{LL}^{(n)})_{max} \\ \vdots \\ V(\hat{\mu}_{LL}^{(n)})_{d_f * d_m} + W_{(d_f * d_m,max)}^{(n)} * V(\hat{\mu}_{LL}^{(n)})_{max} \end{pmatrix} \quad (4.25)$$

Substituting the definition of $W_{(j,max)}^{(n)} = \frac{V(\mu)_j - V(\hat{\mu}_{LL}^{(n)})_j}{V(\hat{\mu}_{LL}^{(n)})_{max}}$ we get $(W^{(n)} + I) * V(\hat{\mu}_{LL}^{(n)}) = V(\mu)$, so we know that $(W^{(n)} + I) * V(\hat{\mu}_{LL}^{(n)})$ satisfies the constraints. Thus, $W^{(n)}$ is a valid solution to our linear programming problem, but it may not minimize our objective function, i.e. $B^{(n)} \neq W^{(n)}$. We are not able to construct $W^{(n)}$ explicitly, because we do not know μ, but we know that such a matrix exists.

$$P(|B_{(i,j)}^{(n)}| \geq 2 * \epsilon) \leq P(\sum_i \sum_j |B_{(i,j)}^{(n)}| \geq 2 * \epsilon) \quad (4.26)$$

$$\leq P(\sum_i \sum_j |B_{(i,j)}^{(n)}| * (1 + V(\hat{\mu}_{LL}^{(n)})_j) \geq 2 * \epsilon) \quad (4.27)$$

$$\leq P(\sum_i \sum_j |W_{(i,j)}^{(n)}| * (1 + V(\hat{\mu}_{LL}^{(n)})_j) \geq 2 * \epsilon) \quad (4.28)$$

$$\leq P(\sum_i \sum_j |W_{(i,j)}^{(n)}| \geq \epsilon) \quad (4.29)$$

$$= P(\sum_i |\frac{V(\mu)_i - V(\hat{\mu}_{LL}^{(n)})_i}{V(\hat{\mu}_{LL}^{(n)})_{max}}| \geq \epsilon) \quad (4.30)$$

The third inequality is true because $B^{(n)}$ minimizes our objective function, and $\sum_i \sum_j |B_{(i,j)}^{(n)}| * (1 + V(\hat{\mu}_{LL}^{(n)})_j)$ is our objective function. The fourth inequality is true because $V(\hat{\mu}_{LL})_j \leq 1$. Since $V(\hat{\mu}_{LL}^{(n)})_i$ is a consistent estimator for $V(\mu)_i$ we get $\lim_{n\to\infty} P(|\frac{V(\mu)_i - V(\hat{\mu}_{LL}^{(n)})_i}{V(\hat{\mu}_{LL}^{(n)})_{max}}| \geq \epsilon) \to 0$ and by Slutsky's Theorem $\lim_{n\to\infty} P(\sum_i |\frac{V(\mu)_i - V(\hat{\mu}_{LL}^{(n)})_i}{V(\hat{\mu}_{LL}^{(n)})_{max}}| \geq \epsilon) \to 0$. Thus, $\lim_{n\to\infty} P(B_{(i,j)}^{(n)} \geq \epsilon) \to 0$. ■

References

Amanatidis, Y., Green, B. and Mihail, M. (2008). Graphic realizations of joint-degree matrices. *Unpublished* .

Bansal, S., Pourbohloul, B. and Meyers, L. A. (2006). A comparative analysis of influenza vaccination programs. *PLoS Medicine* **3** 387.

Barabasi, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512.

Bárány, I. and Füredi, Z. (1987). Computing the volume is difficult. *Discrete and Computational Geometry* **2** 319–326.

Blitzstein, J. and Diaconis, P. (2010). A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Mathematics* **6** 487–520.

Bollobás, B. (2001). *Random Graphs*. 2nd ed. Cambridge University Press, New York.

Chung, F. and Lu, L. (2002). Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics* **6** 125–145. 10.1007/PL00012580.
URL `http://dx.doi.org/10.1007/PL00012580`

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal* **Complex Systems** 1695.
URL `http://igraph.sf.net`

Doherty, I. A., Shiboski, S., Ellen, J. M., Adimora, A. A. and Padian, N. S. (2006). Sexual bridging socially and over time: A simulation model exploring the relative effects of

mixing and concurrency on viral sexually transmitted infection transmission. *Journal of the International AIDS Society* **33** 368–373.

Doyle, J. C., Alderson, D., Li, L., Low, S., Roughan, M., Shalunov, S., Tanaka, R. and Willinger, W. (2005). The 'robust yet fragile' nature of the internet. *Proceedings of the National Academy of Sciences* **102** 14497–14502.

Dyer, M., Friez, A. and Kannan, R. (1991). A random polynomial-time algorithm for approximating the volume of convex bodies. *J. ACM* **38** 1–17.
URL `http://doi.acm.org/10.1145/102782.102783`

Epstein, H. and Stanton, D. (2010). Is polygamy really benign? *AIDS* **24** 1791–1792.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5** 17–61.

Fishburn, P. (1985). *Interval orders and interval graphs: a study of partially ordered sets*. Wiley-Interscience series in discrete mathematics, Wiley.

Fishel, J. D., Ortiz, L. and Barrère, B. (2012). *Measuring concurrent sexual partnerships: Experience of the Measure DHS project to date*. ICF international, Calverton, Maryland, USA. DHS Methodological reports N 7.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association* **81** 832–842.

Friedgut, E. and Kalai, G. (1996). Every monotone graph property has a sharp threshold. *Proceedings of the American Mathematical Society* **124**.

Ghani, A. C. and Garnett, G. (2000). Risks of acquiring and transmitting sexually transmitted diseases in sexual partner networks. *Sexually Transmitted Diseases* **27** 579–587.

Ghani, A. C., Swinton, J. and Garnett, G. (1997). The role of sexual partnership networks in the epidemiology of gonorrhea. *Sexually Transmitted Diseases* **24** 45–56.

Golumbic, M., Kaplan, H. and Shamir, R. (1995). Graph sandwich problems. *Journal of Algorithms* **19** 449–473.

Goodreau, S. (2011). A decade of modelling research yields considerable evidence for the importance of concurrency: a response to sawers and stillwaggon. *Journal of the International AIDS Society* **14** 12.

Goyal, R., Wang, R. and DeGruttola, V. (2012). Editorial commentary: Network epidemic models: Assumptions and interpretations. *Clinical Infectious Diseases* **55** 276–278.

Hakimi, S. (1962). On realizability of a set of integers as degrees of the vertices of a linear graph. *Journal of the Society for Industrial and Applied Mathematics* **10** 496–506.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N. and Morris, M. (2012). *ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks*. Seattle, WA. Version 3.0-1. Project home page at urlstatnet.org. URL `CRAN.R-project.org/package=ergm`

Handcock, M. S. and Jones, J. H. (2004). Likelihood-based inference for stochastic models of sexual network formation. *Theoretical Population Biology* **65** 413Ű422.

Harary, F. (1969). *Graph Theory*. Addison-Wesley, Reading, MA.

Harris, K. M. and Udry, J. R. (2012). *National Longitudinal Study of Adolescent Health (Add Health), 1994-2008*. doi:10.3886/icpsr21600-v9 ed. Inter-university Consortium for Political and Social Research, Ann Arbor, MI.

Havel, V. (1955). A remark on the existence of finite graphs. *Časopis Pest. Mat.* **80** 477–480.

Helleringer, S. and Kohler, H.-P. (2007). Sexual network structure and the spread of hiv in africa: evidence from likoma island, malawi. *AIDS* **21** 2323–2332.

Jeffrey W. Eaton, T. B. H. and Garnett, G. P. (2011). Concurrent sexual partnerships and primary hiv infection: A critical interaction. *AIDS and Behavior* **15** 687–692.

Konde-Lule, J. K., Wawer, M. J., Sewankambo, N. K., Serwadda, D., Kelly, R., Li, C., Gray, R. H. and Kigongo, D. (1997). Adolescents, sexual behavior and hiv-1 in rural rakai district, uganda. *AIDS* **11** 791–799.

Kretzschmar, M. and Morris, M. (1996). Measures of concurrency in networks and the spread of infectious disease. *Mathematical Biosciences* **133** 165–195.

Kretzschmar, M., White, R. G. and Caraël, M. (2010). Concurrency is more complex than it seems. *AIDS* **24** 313–315.

Li, L., Alderson, D., Willinger, W. and Doyle, J. C. (2004). A first-principles approach to understanding the internet's router-level topology. *ACM SIGCOMM Computer Communication Review* **34**.

Lurie, M. and Rosenthal, S. (2010). Concurrent partnerships as a driver of the hiv epidemic in sub-saharan africa? the evidence is limited. *AIDS and Behavior* **14** 17–24. 10.1007/s10461-009-9583-5.
URL `http://dx.doi.org/10.1007/s10461-009-9583-5`

Mahadevan, P., Krioukov, D., Fall, K. and Vahdat, A. (2006). Systematic topology analysis and generation using degree correlations. *Proceedings of the ACM SIGCOMM* .

Maslov, S. and Sneppen, K. (2002). Specificiy and stability in topology of protein networks. *Science* **296** 910–913.

Meyers, L. A., Newman, M. E. J., Martin, M. and Schrag, S. (2003). Applying network theory to epidemics: Control measures for mycoplasma pneumoniae outbreaks. *Emerging Infectious Diseases* **9** 204.

Mills, H. L., Cohen, T. and Colijn, C. (2011). Modelling the performance of isoniazid preventive therapy for reducing tuberculosis in hiv endemic settings: the effects of network structure. *Journal of the Royal Society Interface* **8** 1510–1520.

Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* **6** 161Ű180.

Morris, M. (2010). Barking up the wrong evidence tree. comment on lurie and rosenthal, Şconcurrent partnerships as a driver of the hiv epidemic in sub-saharan africa? the evidence is limitedŤ. *AIDS and Behavior* **14** 31–33. 10.1007/s10461-009-9639-6.
URL `http://dx.doi.org/10.1007/s10461-009-9639-6`

Morris, M., Goodreau, S. and Moody, J. (2007). Sexual networks, concurrency, and std/hiv. In *Sexually Transmitted Diseases* (K. Holmes, S. PF and S. WE, eds.). McGraw-Hill International Book Co, New York, NY, USA.

Morris, M. and Kretzschmar, M. (1997). Concurrent partnerships and the spread of hiv. *AIDS* **11** 641–648.

Morris, M., Kurth, A., Hamilton, D., Moody, J. and Wakefield, S. (2009). Concurrent partnerships and hiv prevalence disparities by race: Linking science and public health practice. *American Journal of Public Health* **99** 1023–1031.

Newman, M. (2002). Assortative mixing in networks. *Physical Review Letters* **89** 208701.

Newman, M. E. (2010). *Networks An Introduction*. Oxford University Press, New York.

Overton, W. S. and Stehman, S. (1995). The horvitzŰthompson theorem as a unifying perspective for probability sampling: with examples from natural resource sampling. *The American Statistician* **49** 261–268.

Palombi, L., Bernava, G. M., Nucita, A., Giglio, P., Liotta, G., Nielsen-Saines, K., Orlando, S., Mancinelli, S., Buonomo, E., Scarcella, P., Altan, A. M. D., Guidotti, G., Ceffa, S., Haswell, J., Zimba, I., Magid, N. A. and Marazzi, M. C. (2012). Predicting trends in hiv-1 sexual transmission in sub-saharan africa through the drug resource enhancement against aids and malnutrition model: Antiretrovirals for reduction of population infectivity, incidence and prevalence at the district level. *Clinical Infectious Diseases* .

Pe'er, I. and Shamir, R. (1995). Interval graphs with side (and size) constraints. In *In Proc. of the Third Annual European Symp. on Algorithms, (ESA 95) Corfu, Greece*. Springer.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org

Reniers, G. and Watkins, S. (2010). Polygyny and hiv: a case of benign concurrency. *AIDS* **24** 299–307.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, New York.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics* **27** 832–837.

Sawers, L., Isaac, A. and Stillwaggon, E. (2011). Hiv and concurrent sexual partnerships: modelling the role of coital dilution. *Journal of the International AIDS Society* **14**. URL http://www.jiasociety.org/index.php/jias/article/view/17472

Sawers, L. and Stillwaggon, E. (2010). Concurrent sexual partnerships do not explain the hiv epidemics in africa: a systematic review of the evidence. *Journal of the International AIDS Society* **13** 34.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York.

Smith, R. L. (1984). Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research* **32** 1296–1308.

Tanser, F., Bärnighausen, T., Hund, L., Garnett, G. P., McGrath, N. and Newell, M.-L. (2011). Effect of concurrent sexual partnerships on rate of new hiv infections in a high-prevalence, rural south african population: a cohort study. *The Lancet* **378** 247–255.

Vázquez, A., Pastor-Satorras, R. and Vespignani, A. (2002). Large-scale topological and dynamical properties of the internet. *Physical Review E* **65** 66130.

Wylie, J. and Jolly, A. (2001). Patterns of chlamydia and gonorrhea infection in sexual networks in manitoba, canada. *Sexually Transmitted Diseases* **28** 14–24.